An Investigation of Dynamic Covariate Effects in Survival Data

Denise Brown

A Dissertation Submitted to the

Faculty of Information and Mathematical Sciences

at the University of Glasgow

for the degree of

Doctor of Philosophy

Department of Statistics

June, 2005

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Survival data are often modelled by the Cox proportional hazards model, which assumes that covariate effects are constant over time. Estimation of covariate effects in such models is usually based on the partial likelihood function with the baseline hazard being estimated non-parametrically. In recent years however, several new approaches have been suggested which allow survival data to be modelled more realistically by allowing the covariate effects to vary with time. Non-proportional hazard functions, with covariate effects changing dynamically, can be fitted using penalised splines (P-splines). Links exist between P-spline smoothing and penalised quasi-likelihood estimation in generalised linear mixed models allowing estimation of the smoothing parameters steering the amount of smoothing. Here a hybrid form for smoothing parameter selection is suggested which combines the mixed model approach with a classical Akaike criterion. Two approaches to estimation of dynamic covariate effects in survival data are considered. One is a Poisson type approach based on the likelihood function and allows for estimation of the baseline hazard, usually treated as a nuisance parameter. The second is a numerically faster approach based on the partial likelihood function. Both approaches are evaluated with simulations and applied to data from the German Socio-Economic Panel. The partial likelihood approach is also applied to data from the West of Scotland Coronary Prevention Study.

Acknowledgements

I would like to thank Professor Göran Kauermann and Professor Ian Ford, my supervisors, for their support and guidance during the course of this research.

Thanks also to the University of Glasgow and the Engineering and Physical Sciences Research Council for providing resources and funding throughout this PhD and to all the staff and postgraduate students in the department of Statistics for making it a welcoming and friendly place to work.

Thanks to Richard Langfield and his family, Katya Kaval, Helen Gilmour and Kate Moran for their support and encouragement. Finally, special thanks to my mum and dad, Frank, Scott, Mark and Tara for everything.

Contents

List of Figures

List of Tables

Chapter 1

Overview

1.1 Aim

The aim of this thesis is to investigate dynamic covariate effects in survival data. Survival data are often modelled using the Cox proportional hazards model. This model assumes that the effect of a covariate measured at the beginning of a study remains constant throughout the duration of the study. However in some cases, particularly those involving long-term follow-up, this assumption of proportional hazards may be unreasonable. It is therefore of interest to be able to model survival data more realistically by allowing covariate effects to vary with time.

1.2 Outline of the Thesis

Chapter 2 begins by giving a brief introduction to survival analysis. Alternative methods of describing the distribution of survival times are discussed, with focus lying on the hazard function which can be modelled either parametrically or non-parametrically. One feature of survival data is that the time to event is not necessarily observed in all subjects. These non-observed events are known as censored observations. When interest lies in exploring the relationship between the survival of a patient and several explanatory variables the Cox proportional hazards model, described in Section 2.4, is often used. This model is considered semi-parametric in that a parametric form is assumed for the effects of explanatory variables, but it allows an unspecified form for the underlying baseline hazard function. An important assumption of the Cox proportional hazards model is that covariate effects are assumed to be constant over time. Covariate effects are estimated using the method of partial likelihood estimation, as described in Section 2.4.2. Maximum likelihood estimates of covariate effects are found by maximising the log partial likelihood function using numerical methods such as the Newton-Raphson procedure. The partial likelihood for the Cox proportional hazards model assumes that there are no ties between event times. Approaches for constructing the partial likelihood when there are ties among the event times are given in 2.4.4. In 2.4.5 the interpretation of covariate effect estimates are described in the case of both continuous and categorical covariates. Once estimated, the covariate effects allow inferences to be made about the effect of explanatory variables on the hazard function. However in order to obtain estimates of the hazard function for an individual, an estimate of the baseline hazard function is required. Methods for estimating the baseline hazard function are discussed in

Section 2.5. The Cox proportional hazards model can be extended to include time-varying covariates. Here, baseline covariate values are updated over the follow-up period of study so that covariate values in the Cox model now depend on time. A further extension of the Cox model can be made by allowing the *effects* of covariates to vary with time. This enables one to see whether baseline covariates become more or less prognostic with time. Interest lies in modelling these dynamic covariate effects which vary smoothly with time.

In *Chapter 3* an overview of penalised spline smoothing (P-splines) is given. The Chapter begins by describing some of the simple parametric methods used to estimate relationships in regression analyses, such as linear regression or linear interpolation. Polynomial regression extends simple linear regression and can be used to handle nonlinear structures in the data, however it is most suitable when the pattern of nonlinearity is fairly simple. To increase flexibility, scatterplot smoothers, such as local smoothers, are used. Although flexible, local smoothers can be slow computationally. Spline-based smoothers are another approach to smoothing. Section 3.4 describes regression splines and the problems concerned with knot selection. In Section 3.5, P-splines are introduced and the question of selecting appropriate smoothing parameters now arises. A link between P-spline smoothing and generalised linear mixed models exists such that a data driven estimate of smoothing parameters can be obtained. The link is illustrated for normal responses. The Chapter ends with an overview of smoothing in survival models and motivates the use of P-splines for smooth hazard modelling.

The first of two approaches used to estimate dynamic effects in survival data is detailed in *Chapter 4*. Usually covariate effects in the Cox proportional hazards model are estimated via the partial likelihood function. However, there are two

main reasons why it makes sense to work directly with the likelihood function. These reasons are described in Section 4.2. In Section 4.3 the standard Cox model is rewritten in order to include dynamic covariate effects and an estimate of the baseline hazard function. The penalised likelihood is constructed, containing integrals based on the hazard function. As a result numerical integration is employed in Section 4.4 leading to a Poisson type model. In Section 4.5 the relationship between P-splines and generalised linear mixed models is considered again, this time in the presence of non-normal response models. This link is used for smoothing parameter selection. The smoothing parameter estimate tends to over smooth the data hence a hybrid smoothing parameter estimate, controlled by the Akaike criterion, is suggested in 4.5.2. Finally simulations showing the performance of this procedure ends the Chapter.

The second approach used to estimate dynamic covariate effects in survival data is described in *Chapter 5*. This approach, based on the partial likelihood function, does not explicitly estimate the baseline hazard function however it is numerically faster than the Poisson type approach described in Chapter 4. This enables estimation of covariate effects in large survival data sets. In Section 5.2 the penalised partial log likelihood is constructed. Estimation of smooth covariate effects is achieved using P-splines and the link between P-splines and generalised linear mixed models is utilised for smoothing parameter selection. A hybrid smoothing parameter estimate controlled by the Akaike criterion is obtained. Simulations end this Chapter.

In *Chapter 6* two data sets are introduced. The first is unemployment data from the German Socio-Economic Panel Study. Both the Poisson type approach from Chapter 4 and the partial likelihood approach from Chapter 5 are used to

investigate dynamic covariate effects in this study. The second data set is based on the West of Scotland Coronary Prevention Study (WOSCOPS) and is analysed using the partial likelihood approach from Chapter 5.

The final Chapter gives a summary of the work completed and conclusions drawn.

Chapter 2

An Introduction to Survival Analysis

2.1 Introduction

Survival analysis involves studying the time taken until a particular event occurs. This type of analysis is used in many fields including social science, industrial reliability and in medical studies where the event of interest is often the death of a patient. It could also be the time to response associated with a treatment or the time to development of a disease so the term survival time could be more correctly defined as the time to event, since not all events involve death. One distinguishing feature of survival data is that the distribution of survival times will tend to be *positively skewed.* As a result, it will not be reasonable to assume that data of this kind will have a normal distribution. Another feature is that the time to event is not necessarily observed in all subjects. These non-observed

events are defined as censored observations.

2.2 Censoring

Censored observations are those observations for which the end-point of interest has not been observed, the term censoring first being used by Hald (1949). In the case of *right-censoring* this could be due to the fact that an individual has been lost to follow-up or that an individual has not yet experienced the event of interest by the end of the study. All that is known about the observation is that the event of interest occurs later than some given point in time.

In contrast, *left-censoring* occurs when an individual has experienced the event of interest at some time prior to the beginning of the study. For example, one may know that a patient entered hospital on a particular date, and that the patient survived for a certain amount of time thereafter; however, it may not be known exactly when the symptoms of the disease first occurred.

Another form of censoring is *interval-censoring*. This type of censoring occurs when the failure time is known to occur only within an interval. Usually this data comes from a trial where the objects of interest are not constantly monitored. For example, in a clinical trial an event may occur at some unknown point between two clinical examinations. Examples of censoring are given in Klein and Moeschberger (1997).

In the case of right-censoring, a distinction can be made between three different types of censoring. With *Type I* censoring, the subjects enter the study at the

same time with the event observed only if it occurs prior to some pre-specified time end-point. Subjects also enter the study at the same time when *Type II* censoring is used but here the end of the study is not initially fixed. Instead the study continues until a predetermined number of subjects have experienced an event. Alternatively with *Type III* censoring, subjects enter the study at different times.

2.3 Distribution of Survival Times

There are three alternative ways for describing the distribution of survival times, namely the *probability density function*, the *survival function* and the *hazard function*. Let t be the actual survival time of an individual, regarded as the value of a variable T, which can take any non-negative value. Then T is the random variable associated with the survival time with distribution function $F(t)$ and density function $f(t)$. The distribution function of T is given by

$$F(t) = P(T < t) = \int_0^t f(u)du,$$

and represents the probability that survival is less than some time t.

The survival function $S(t)$, represents the probability of surviving at least as long as t, which is 1 minus the distribution function of T, i.e

$$S(t) = P(T \geq t) = 1 - F(t).$$

It can be used to represent the probability that an individual survives from the

time origin to some time beyond t.

Finally, the hazard function is the probability of failing in the next small interval δ_t having already survived to time t. The hazard function is written as

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta_t | T \geq t)}{\delta_t} \right\},$$

and is defined as the instantaneous rate of failure. The cumulative hazard function $H(t)$ can be expressed as

$$H(t) = \int_0^t h(u)du,$$

and can be interpreted as the probability of failing at time t having survived until that time. These functions are used to illustrate different aspects of the distribution of T and are mathematically equivalent to each other. So given one, the other functions can be uniquely determined.

The survival times can be modelled *parametrically* or *non-parametrically.* Commonly used parametric distributions are the exponential and Weibull distributions. Here, methods for estimation are based on the assumption of a particular form of probability distribution for the survival times. If these assumptions for the data are valid, inferences will be more precise than they would in the absence of a distributional assumption.

The simplest parametric model for the survival times is to assume that it comes from an exponential distribution, characterised by one parameter, λ. Under this model, the hazard function may be written as

$$h(t) = \lambda,$$

where λ is a positive constant. This constant hazard rate is a unique property of the exponential distribution and implies that the probability of failure remains constant over time, i.e. the hazard rate does not depend on time.

However, in many cases this assumption of a constant hazard rate will not hold. For example, following major surgery, death is more likely to occur immediately after an operation with the level of risk reducing thereafter. One way of making the hazard depend on time is to assume that the survival times come from a Weibull distribution. Under a Weibull distribution, the hazard function is

$$h(t) = \lambda\gamma t^{\gamma-1},$$

a function which depends on two parameters λ and γ, both greater than zero. The parameter γ is known as the shape parameter, and λ as the scale parameter. In the particular case where $\gamma = 1$, the hazard function has a constant value λ, and the survival times have an exponential distribution. Since the Weibull hazard function can take a variety of forms depending on the value of the shape parameter γ, this distribution is used widely in the parametric analysis of survival data.

Methods which do not require the form of the probability density function of T to be specified are said to be non-parametric or distribution-free. In the analysis of a single sample of survival data the life table estimate of the survival times is one useful non-parametric method used, another is the Kaplan-Meier estimate (Kaplan and Meier, 1958). A simple method of estimating the hazard function when there is a single sample of survival data is to take the ratio of the number of events at a given time to the number of subjects at risk at that

time point. Two non-parametric procedures for comparing two or more groups of survival times are the log-rank test (Mantel and Haenszel, 1959) and the Wilcoxon test (Gehan, 1965). An introduction to these procedures can be found in Collett (1994).

2.3.1 The Effect of Censoring on the Likelihood

The probability density function is denoted by $f(t;\gamma)$ where γ represents the unknown parameters. Assuming that the survival times are independent, the likelihood function for observed survival times can be written as

$$L(\gamma) = \prod_{i \in \mathcal{D}} f(t_i;\gamma),$$

where $\mathcal{D}$ is the set of units failing. However, this likelihood fails to take into account those observations which have been censored. Recall that the survival function $S(t)$ gives the probability of surviving to time t. In the case of right-censoring, all that is known is that the subject survived until a particular time point. Let $S(R_i;\gamma)$ be the contribution to the likelihood from a right-censored observation. Denote by $1-S(L_i;\gamma)$ the contribution to the likelihood from a left-censored observation. Finally let $[S(L_i;\gamma) - S(R_i;\gamma)]$ be the contribution to the likelihood from an interval-censored observation. Then the likelihood function is simply the joint probability of the uncensored and censored observations

$$L(\gamma) = \prod_{i \in \mathcal{D}} f(t_i;\gamma) \prod_{i \in \mathcal{R}} S(R_i;\gamma) \prod_{i \in \mathcal{L}} (1 - S(L_i;\gamma)) \prod_{i \in \mathcal{I}} [S(L_i;\gamma) - S(R_i;\gamma)],$$

where $\mathcal{R}$ is the set of right-censored observations, $\mathcal{L}$ the set of left-censored observations and $\mathcal{I}$ the set of interval-censored observations.

2.4 Cox Proportional Hazards Model

The Cox proportional hazards model (Cox, 1972) explores the relationship between the survival of a patient and several explanatory variables. This model is considered *semi-parametric* in that a parametric form is assumed for the effects of the explanatory variables, but it allows an unspecified form for the underlying hazard function.

In the Cox proportional hazards model, the conditional hazard function is assumed to be of the form

$$h(t|\beta, x_i) = h_0(t) \exp(\beta^T x_i), \tag{2.1}$$

where x_i is a p dimensional set of covariates for the ith individual, $h_0(t)$ is the unspecified baseline hazard function and β is a vector of unknown covariate effects. The hazard function is therefore a product of two functions: the underlying baseline hazard function $h_0(t)$ which describes how the hazard function changes as a function of time, and the risk score which describes how the hazard function changes as a function of the covariates. A key assumption of this model is the proportionality of the hazards, meaning that covariate effects are assumed to be constant over time.

2.4.1 Assessing Proportionality of the Hazards

The Cox model assumes that the ratio of the hazard functions for any two subgroups, i.e. two groups with different values of the covariate x, is constant over time. This assumption of proportional hazards is a strong one and it is important that its appropriateness is checked. In the case of categorical covariates, a plot of the log cumulative hazard versus time (log-log plot) is a standard graphical tool used to indicate a violation of the proportional hazards assumption. The logarithm of the survival time is plotted against the estimated log cumulative hazard ($\log[-\log(s(\hat{t}))]$). If the plotted curves for the different subgroups are approximately parallel then the proportional hazards assumption is justified. Hess (1995) reviews graphical methods for assessing the appropriateness of the proportional hazards assumption. Cox (1972) and Grambsch and Therneau (1994) suggest a parametric extension by including time-varying covariates in the model. Tests of this kind require the pre-specification of a suspected departure from proportionality in a functional parametric form. For a general overview of estimation and tests in proportional hazards models see, for example, Sasieni (1999).

2.4.2 Partial Likelihood Estimation

The regression coefficients β in Equation (2.1) are estimated using partial likelihood estimation. The idea behind using the partial likelihood for estimation is that no information about the effect of the explanatory variables on the hazard function is available from time intervals in which no failures occur.

Assuming no ties, suppose that there are n distinct death times among N

individuals. Death times are denoted by $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$. Let $R(t_{(j)})$ be the set of all individuals at risk at time $t_{(j)}$, $j = 1, \ldots, n$. This is the set of all subjects alive and uncensored just prior to time $t_{(j)}$ and is known as the *risk set.* Let $x_{(j)}$ be the vector of covariates for the individual who dies at the jth ordered death time $t_{(j)}$. Then the probability that an individual with covariates $x_{(j)}$ dies at time $t_{(j)}$, given one of the individuals in $R(t_{(j)})$ dies at that time point, is given by

$$P[\text{individual dies at } t_{(j)} \mid \text{one death at } t_{(j)}]$$

$$= \frac{P[\text{individual dies at } t_{(j)} \mid \text{survival to } t_{(j)}]}{P[\text{one death at } t_{(j)} \mid \text{survival to } t_{(j)}]}.$$

The hazard of death at time $t_{(j)}$ for an individual with covariates $x_{(j)}$ is shown in the numerator above. The expression in the denominator gives $h_{(j)}$, the sum of the values over all individuals, indexed by l, in the risk set at time $t_{(j)}$, $R(t_{(j)})$.

$$= \frac{h[t_{(j)} \mid x_{(j)}]}{\sum_{l \in R(t_{(j)})} h[t_{(j)} \mid x_l]} = \frac{h_0(t_{(j)}) \exp[\beta^T x_{(j)}]}{\sum_{l \in R(t_{(j)})} h_0(t_{(j)}) \exp[\beta^T x_l]}.$$

Cancelling out the baseline hazard function in the numerator and denominator gives

$$= \frac{\exp[\beta^T x_{(j)}]}{\sum_{l \in R(t_{(j)})} \exp[\beta^T x_l]}.$$

By multiplying these conditional probabilities over all individuals for whom death times have been recorded, the partial likelihood is formed

$$L(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T x_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\beta^T x_l)}. \tag{2.2}$$

Here the summation in the denominator of the likelihood is the sum of the values of $\exp(\beta^T x)$ over all individuals at risk at time $t_{(j)}$. Inference about the effect of

explanatory variables on the hazard function depends only on the rank order of survival times since the likelihood function depends only on the ranking of the death times.

Now define δ_i as the censoring indicator for the ith patient where $\delta_i = 1$ if the event has occurred and $\delta_i = 0$ if the ith survival time t_i, $i = 1 < \ldots < N$, is right-censored. The partial likelihood function above can therefore be expressed as

$$L(\beta) = \prod_{i=1}^{N} \left\{ \frac{\exp(\beta^T x_i)}{\sum_{l \in R(t_i)} \exp(\beta^T x_l)} \right\}^{\delta_i},$$

where $R(t_i)$ is the risk set at time t_i. More details on the theory and applications of the proportional hazards model may be found in Cox and Oakes (1984).

2.4.3 Newton-Raphson Procedure

In the Cox proportional hazards model, maximum likelihood estimates of the β parameters can be found by maximising the log-likelihood function using numerical methods such as the *Newton-Raphson Procedure*. The partial log likelihood function $l(\beta) = \log L(\beta)$, can be expressed as follows

$$l(\beta) = \sum_{i=1}^{N} \delta_i (\beta^T x_i) - \sum_{i=1}^{N} \delta_i \log \left\{ \sum_{l \in R(t_i)} \exp(\beta^T x_l) \right\}.$$

Let $U(\beta)$ be the $p \times 1$ vector of first derivatives of the log-likelihood function with respect to the β parameters, known as the *vector of efficient scores*, where p is the number of covariates recorded for each individual. This quantity is given by

$$U(\beta) = \frac{\partial l}{\partial \beta} = \delta^T X - \sum_{i=1}^{N} \frac{\sum_{l \in R(t_i)} exp(\beta^T x_l) X_{(l,)}}{\sum_{l \in R(t_i)} exp(\beta^T x_l)},$$

where $\delta = (\delta_1, \ldots, \delta_N)^T$ is the vector of censoring indicators, and X is the $(N \times p)$ matrix of covariate values, with the covariate values of the lth individual, $X_{(l,)} = x_l^T$, contained in the lth row. Also, let $I(\beta)$ be the $p \times p$ matrix of negative second derivatives of the log-likelihood. This is known as the *observed information matrix*,

$$I(\beta) = -\frac{\partial^2 l}{\partial \beta^2}.$$

The partial maximum likelihood estimates are then found by solving the set of p nonlinear equations $U(\beta) = 0$. Using the Newton-Raphson procedure, an estimate of the vector of β parameters at the $(k+1)$'th cycle, for $k = 0, 1, 2, \ldots$, of the iterative procedure, $\hat{\beta}_{k+1}$, is

$$\hat{\beta}_{k+1} = \hat{\beta}_k + I(\hat{\beta}_k)^{-1} U(\hat{\beta}_k),$$

where $I(\hat{\beta}_k)^{-1}$ is the inverse of the observed information matrix and $U(\hat{\beta}_k)$ is the vector of efficient scores, both evaluated at $\hat{\beta}_k$. Generally, the process can be started by taking $\hat{\beta}_0 = 0$.

Estimated asymptotic variances may be obtained from the inverse of the information matrix.

2.4.4 Handling of Ties

The partial likelihood in Equation (2.2) assumes that there are no ties between the event times. However, often survival times are recorded to the nearest day, week, month or year and this may result in more than one failure occurring at the

same time. When there are tied failure times it becomes unclear which individuals to include in the risk set at each failure time $t_1, t_2, t_3, \ldots$. The exact likelihood function has to include all possible orderings of tied failures (Kalbfleisch and Prentice, 1980), and hence is very difficult computationally. Several approximations to the partial likelihoods have been proposed in the event of ties. Recall that $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$ denote the n distinct, ordered, event times and let $R(t_{(j)})$ be the set of all individuals at risk just prior to $t_{(j)}$. Let d_j be the number of deaths at time $t_{(j)}$ and let $\mathcal{D}_j$ be the set of all individuals who die at time $t_{(j)}$. Now let $s_j = \sum_{l \in \mathcal{D}_j} \mathbf{x_l}$ be the sum of the covariate vectors over all individuals who die at $t_{(j)}$. Breslow's approximation (Breslow, 1974) is

$$L_B(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T s_j)}{\left[\sum_{l \in R(t_{(j)})} \exp(\beta^T x_l)\right]^{d_j}}. \tag{2.3}$$

When there are few ties this method works well. Efron (1977) proposed an approximation which is a slightly better approximation to the exact partial likelihood than Breslow's approximation. The Efron likelihood is

$$L_E(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T s_j)}{\prod_{k=1}^{d_j} \left[\sum_{l \in R(t_{(j)})} \exp(\beta^T x_l) - \frac{k-1}{d_j} \sum_{l \in \mathcal{D}_j} \exp(\beta^T x_l)\right]}. \tag{2.4}$$

When the number of ties is small, Breslow's and Efron's likelihoods are quite close. When $d_j = 1$, the terms in the numerators and denominators of Equations (2.2), (2.3) and (2.4) are identical.

2.4.5 Interpretation of Parameter Estimates

In the Cox proportional hazards model, a single covariate effect β, can be interpreted as the logarithm of the ratio of the hazard of event for a particular individual to the baseline hazard, with the baseline hazard in this case representing the 'average' individual.

In the case of a *continuous* covariate, e.g weight (kg), then the estimated coefficient $\exp(\beta)$ is the estimated change in the hazard ratio when the value of the covariate is increased by 1 unit. If for example, $\exp(\beta)$ is greater than 1, then the rate of experiencing the event of interest increases with each increasing unit. If however $\exp(\beta)$ is less than 1 then the rate of the event decreases with each increasing unit. So, if $\exp(\beta) = 1.2$ for weight in kgs then the rate of experiencing an event increases by 1.2 for every kg increase in weight. This is independent of the weight at which the increase is calculated.

In the case of two *categories*, e.g treatment group, a code must be assigned to each of the possible outcomes. Usually the two levels are coded as 0 and 1 (e.g 0=placebo and 1=treatment), where 0 represents the baseline category. In this case, if $\exp(\beta)$ is greater than 1 then those who have the second level of the covariate (coded 1) are at higher risk while if $\exp(\beta)$ is less than 1, then those in the first level (coded 0) are at higher risk. So, if $\exp(\beta) = 0.8$ for the treatment group then survival is poorer for those who are receiving the placebo. When a categorical covariate has more than two levels it is usual to fit a Cox model using dummy variables, see for example Parmar and Machin (1995).

2.5 Estimating the Baseline Hazard

So far, the focus has been on estimating the covariate effects β. Once estimated, the covariate effects enable inferences to be drawn about the effect of explanatory variables in the model on the hazard function. In order to obtain estimates of the hazard function for an individual, an estimate of the baseline hazard function $h_0(t)$ is also required. Cox (1972) suggested a point-wise estimate for $h_0(t)$ which is identically zero, except at the points where failure occurs. Oakes (1972) instead proposed a step-function estimate, in the discussion that followed Cox's paper, based on the assumption that $h_0(t)$ is a function which varies slowly with time and additionally suggested applying some grouping or smoothing procedure to the estimates in order to obtain a good indication of the behaviour of the baseline hazard function. Another contribution to the discussion of Cox's paper was given by Breslow (1972) who proposed a step-function estimate for $h_0(t)$ where the function $h_0(t)$ is assumed to be constant between those intervals of time in which the event occurs. Kalbfleisch and Prentice (1973) used a similar estimate but suggested that the baseline hazard function is constant between suitable, but arbitrary, time intervals. A piece-wise smooth estimate is given by Anderson and Senthilselvan (1980) which is based on penalised maximum likelihood methods. The advantage of this method is that there is no constraint on the form of the function $h_0(t)$.

2.6 Time-varying Covariates

Covariates are often measured at the beginning of the study. These measurements are known as *baseline* values. However, individuals can be monitored for the entire duration of the study. Measurements which are updated during the follow-up period may be more predictive of survival experience than the original baseline values. There are two types of variable that change over time, these are referred to as *internal variables* and *external variables.* Internal variables are variables which are generated by the individual under study and are only observed as long as the patient survives. An example of this is blood pressure, which may be assessed at entry into a study but has the potential to change thereafter. In contrast, an external variable is one whose value at a particular time does not require individuals to be under direct observation. One type of external variable is age which changes in such a way that its value will be known in advance at any future time point. If patients are followed over a long period of time, their current age may be more predictive of survival than their age when the study began. These types of covariates can be introduced into the Cox model. Generalising to include time-varying covariates, the model in Equation (2.1) becomes

$$h(t|\beta, x_i) = h_0(t) \exp(\beta^T x_i(t)).$$

In this model, the values of the covariates depend on time t. Therefore the hazard of death at time t is no longer proportional to the baseline hazard and the model is no longer a proportional hazards model. Instead it is referred to as the Cox regression model. The possibility of incorporating such variables in a proportional hazards model was first explored by Cox (1972) and the appropriate

partial likelihood function is discussed in Cox (1975).

2.7 Dynamic Covariate Effects

In most studies involving survival data, proportional hazards are assumed for the covariate effects. This means that the effect of a variable measured at the beginning of a study remains constant throughout the duration of the study. However in some cases, especially those involving long-term follow-up, this assumption may not be reasonable. In breast cancer studies for example, the size of a tumour may strongly influence the short term prognosis, but may not be relevant after a patient has remained disease-free for some time. In this case a *dynamic* effect may exist. This idea is different to that in the previous Section. Here it is the *effects* of the covariates that are changing with time and not the actual covariate values themselves. The Cox model may be extended to include dynamic covariate effects

$$h(t|\beta, x_i) = h_0(t) \exp(\beta^T(t) x_i),$$

as generally introduced by Hastie and Tibshirani (1993), where $\beta(t)$ is a vector of covariate effects varying smoothly with time t.

2.8 Chapter Summary

Survival analysis studies time to event data. A distinguishing feature of this type of data is the censored values which may be observed, another is the fact that distributions are often positively skewed. Survival data can be modelled using the

Cox proportional hazards model where the main assumption is proportionality of the hazards. This means that covariate effects are assumed to remain constant over time. The Cox model can be extended to include dynamic covariate effects which vary smoothly in time. The main focus in the following Chapter is to motivate the use of penalised splines for the smooth modelling of dynamic covariate effects.

Chapter 3

An Introduction to Penalised Spline Smoothing

3.1 Introduction

Traditional parametric regression methods often fail to capture important details present in real data. However there is a need to be able to handle complex relationships effectively by using more flexible techniques. Smoothing methods exist which aim to provide a means of modelling such data. In this Chapter, the main ideas of smoothing are introduced and the use of penalised spline (P-spline) smoothing is motivated. In addition the link between penalised spline smoothing and generalised linear mixed models will be highlighted allowing appropriate estimation of smoothing parameters.

3.2 Parametric Approach

Suppose that responses $y_1, \ldots, y_n$ observed at design points $x_1 < \ldots < x_n$ follow the regression model

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n. \tag{3.1}$$

The function f is an unspecified function to be estimated from the (x_i, y_i) and $\epsilon_1, \ldots, \epsilon_n$ are zero-mean, uncorrelated random errors. There are several methods available for estimating the function f. Consider a simulated example, with the sample generated according to the relation

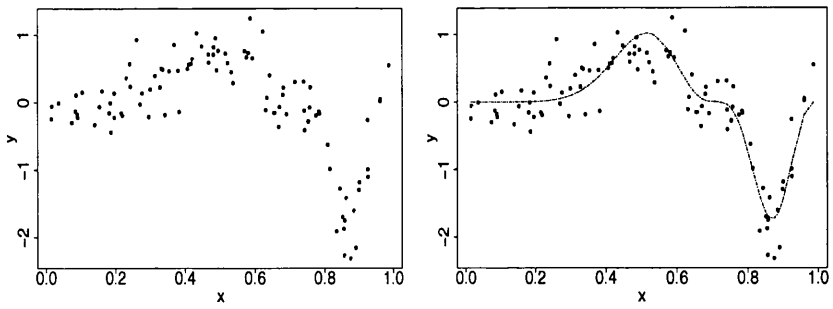
$$y_i = 2x_i^2 sin^3(2\pi x_i^2) + \epsilon_i, \qquad i = 1, \ldots, 100, \tag{3.2}$$

where the x_i's are drawn from the uniform distribution $U(0, 1)$ and the ϵ_i's are drawn from the normal distribution $N(0, 0.3)$. Figure 3.1(a) shows the graphical representation of this model with the true underlying regression line shown in Figure 3.1(b).

One of the most common ways to estimate f is by simple linear regression. Using this technique, the function $f(x)$ is estimated by $\hat{a} + \hat{b}x$ where $\hat{a}$ is the least squares intercept estimator and $\hat{b}$ the least squares slope estimator. These are obtained by minimising the *residual sum of squares*

$$RSS = \sum_{i=1}^{n} \{y_i - \hat{f}(x_i)\}^2. \tag{3.3}$$

Figure 3.2 shows the linear regression fit to the data simulated in Equation (3.2).



(a) (b)

Figure 3.1: *(a) The simulated data; (b) the true regression line.*

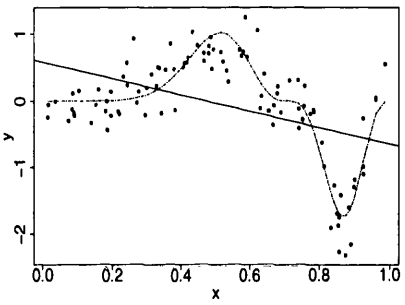


Figure 3.2: *Simple linear regression fit to the simulated data. The true function is given by the dashed line.*

This approach to estimation can be very effective if the underlying regression function f in Equation (3.1) is approximately linear. However, in cases where a more complex relationship exists linear regression tends to smooth out important features of the data. Another approach is to estimate f by linear interpolation. Linear interpolation is used to estimate values of a function between two known values. Unlike the linear regression fit which uses too little of the data, linear interpolation uses too much information. It summarises the random noise component of the model rather than providing a useful description of the true regression function. Obviously, some compromise between the two approaches is needed. Polynomial regression is an extension of simple linear regression and

can be used to handle nonlinear structure in the data. Polynomial regression fits data to the equation: $f(x) = a + bx + cx^2 + dx^3 + \ldots$ where any number of terms can be included. Stopping at the second term results in the equation for a straight line and is a first-order polynomial. Stopping after the third term results in a quadratic equation, or a second-order polynomial and so on. Figure

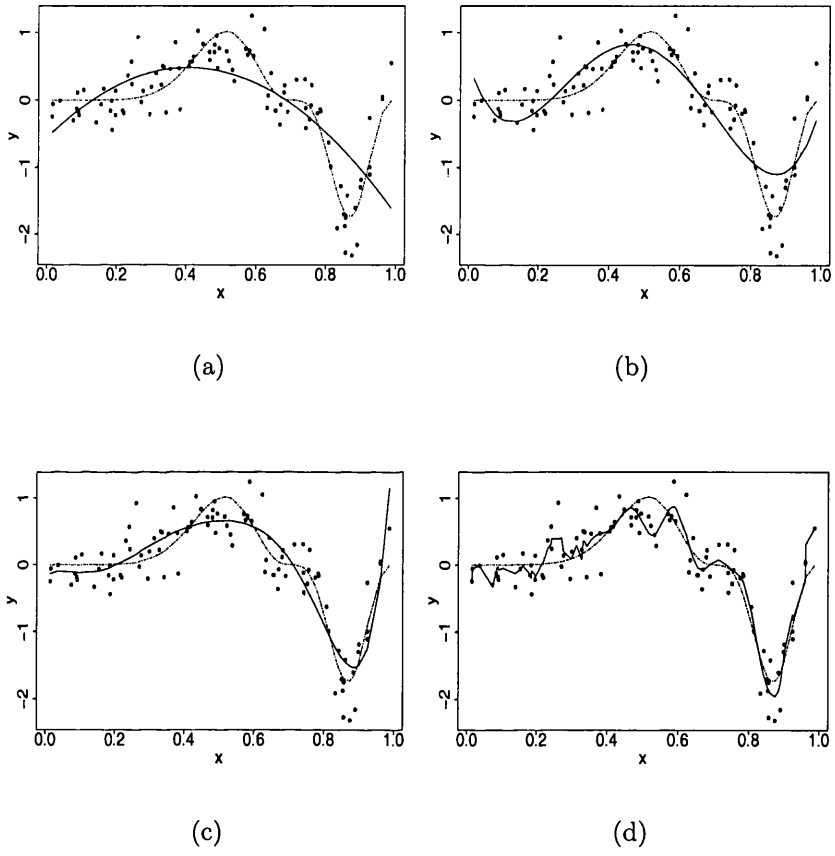


Figure 3.3: *Polynomial regression fits to the simulated data of degree (a) 2; (b) 4; (c) 7; (d) 30. The true function is given by the dashed line.*

3.3 shows polynomial regression fits to the data, of varying degrees. In Figure 3.3(a), a polynomial of degree 2 is used. This quadratic fit over smooths the data, so it is worth choosing a polynomial of higher degree in order to try to capture the curvature in the data. Figure 3.3(b) shows the fit for a polynomial

of degree 4. This model fits the data better but fails to reach the peaks as it over smooths the data. A polynomial of degree 7, in Figure 3.3(c) captures more of the trend at the peaks, but still over smooths the data. As the degree of the polynomial increases, see Figure 3.3(d), the estimator ends up interpolating the observed points. Polynomials are often used when a simple empirical model is required, however as the degree of the polynomial increases the model becomes hard to interpret without a graph. This suggests that polynomial regression is most suitable when the pattern of nonlinearity is fairly simple, however other approaches are needed when the nonlinearity is more complex.

3.3 Local Smoothing

The parametric approaches described above are not flexible enough to capture nonlinear effects. A *smoother*, is a tool for estimating the trend of a response measurement. An important property of a smoother is that it is nonparametric in nature and hence avoids the assumption of a rigid form for the dependence on the x_i's. In the case of a single predictor, it is common to use the term *scatterplot smoothing* as interest lies in describing the underlying trend in the scatterplot.

One method of scatterplot smoothing is *local smoothing*. A set of local weights, defined by a *kernel*, are used to produce estimates at values of x. Various kernels can be used, a popular choice of kernel function K, is the Gaussian (bell-shaped) kernel. The Gaussian kernel smoother uses the Gaussian density function to assign weights to neighbouring points where the largest weight is assigned to the target point x and weights decrease symmetrically as one moves away from this

point. Typically the kernel smooth is computed as the local mean estimator

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)},$$

first proposed by Nadaraya (1964) and Watson (1964). The parameter h is a *smoothing parameter*, or *bandwidth*, controlling the width of the kernel function and hence the amount of smoothing. For a Gaussian kernel function h is its standard deviation. Figure 3.4 illustrates the effects of using different bandwidths

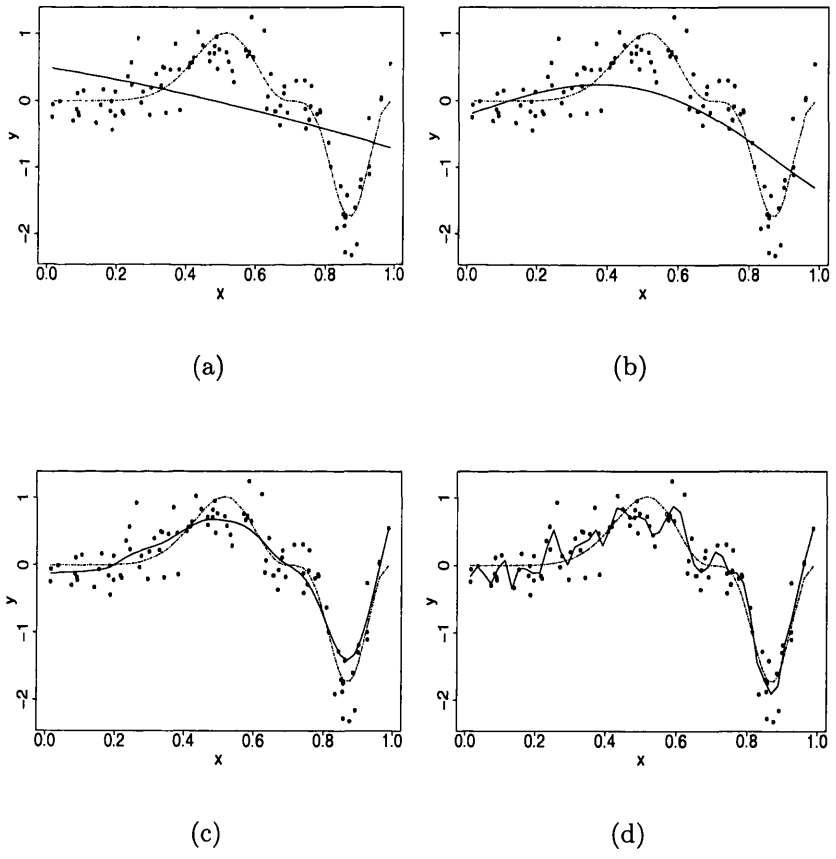


Figure 3.4: *Kernel smoothing fits to the simulated data with bandwidths (a)* h = 1*; (b)* h = 0.25*; (c)* h = 0.05*; (d)* h = 0.01*. The true function is given by the dashed line.*

h. For larger values of h the estimator misses some curvature in the data while

as the smoothing parameter h decreases, the estimator begins to interpolate the observed points. Various smoothers, including local smoothers, are described in Hastie and Tibshirani (1993) while Bowman and Azzalini (1997) show how these smoothing techniques are used in practice.

3.4 Regression Splines

Computationally, local smoothers can be slow. *Spline-based smoothers* offer another approach to scatterplot smoothing. *Splines* are *piecewise polynomial functions* that are constrained to join at certain values of x called the *knots*. Consider the linear regression model $f(x) = a + b_1 x$. The term $a + b_1 x$ is a *linear combination* of the *basis functions* 1 and x. These basis functions correspond to the columns of the design matrix in regression

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_1 \end{bmatrix}$$

In polynomial regression, there are extra basis functions which correspond to the extra terms in the model. Therefore the basis matrix for the cubic model is

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$$

Now, consider a *linear spline regression* with two knots t_1 and t_2 given by

$$f(x) = a + b_1 x + \eta_1 (x - t_1)_+ + \eta_2 (x - t_2)_+,$$

where $(x - t_1)_+$, for example, is equal to $x - t_1$ if $x - t_1$ is positive and is equal to 0 otherwise. For this regression model the basis matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1 - t_1)_+ & (x_1 - t_2)_+ \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - t_1)_+ & (x_n - t_2)_+ \end{bmatrix}$$

More functions of the form $(x-t)_+$ can be added to the basis. Let q be the number of knots. Thus the spline model for linear basis functions $1, x, (x - t_1)_+, \ldots, (x - t_q)_+$ is

$$f(x) = a + b_1 x + \sum_{\kappa=1}^{q} \eta_\kappa (x - t_\kappa)_+.$$

3.4.1 Cubic Splines

One of the most commonly used piecewise polynomials in spline regression is the *cubic spline*, which is constrained to be continuous and have continuous first and second derivatives at the knots. A cubic regression spline with two knots t_1 and t_2 can be written as

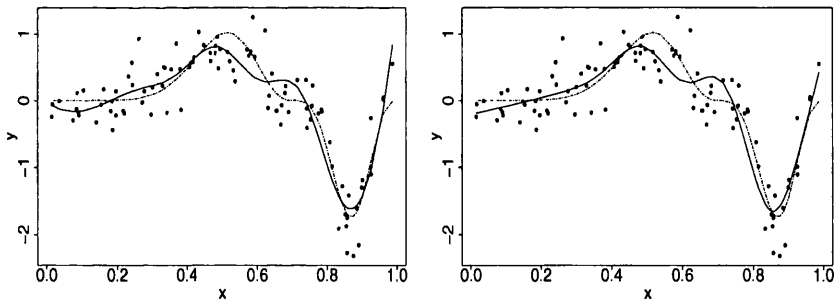
$$f(x) = a + b_1 x + b_2 x^2 + b_3 x^3 + \eta_1 (x - t_1)_+^3 + \eta_2 (x - t_2)_+^3.$$

In general if there are q knots, the function will require $q+4$ regression coefficients. *Natural cubic splines* extend the cubic spline by adding knots at the boundaries

of the data, thus constraining the fit to be linear after the boundary knots. To enforce this condition, the requirement that the first and second derivatives are continuous at t_1 and t_q is dropped. A natural spline can be written as

$$f(x) = a + b_1 x + \eta_1 (x - t_1)_+^3 + \eta_2 (x - t_2)_+^3, \ldots, + \eta_q (x - t_q)_+^3,$$

and requires $q + 2$ regression coefficients. Figure 3.5(a) shows the fit of a cubic spline and Figure 3.5(b) shows the natural spline fit having the constraint that the function is linear beyond the boundary knots. While regular cubic splines can have high variance near the boundary, natural splines are less flexible at the boundaries as a result of the linear constraint. However there is little reason for preferring the natural cubic spline model to the cubic spline model (Ruppert, Wand and Carroll, 2003).



(a) (b)

Figure 3.5: *(a) A cubic spline fit to the simulated data; (b) a natural spline fit to the simulated data. The true function is given by the dashed line.*

3.4.2 Knot Selection

Natural splines, and cubic splines, are *fixed-knot* splines. Therefore the number of knots need to be selected as well as the placement of the knots. Stone (1986) found that it matters less where the knots are than how many knots are chosen, especially in the case of natural splines. Placing knots at equally spaced intervals on the x range, ensures that there is enough data within each interval to get a sensible fit and also guards against outliers overly influencing the curve. In Figure 3.5, the number of knots ($q = 6$) were chosen subjectively. Choosing too few knots results in less of the curvature being captured, while choosing too many knots results in overfitting. *Akaike's Information Criterion* (AIC) (Akaike, 1973) can be used for a data driven selection of q where the value of q chosen is that which gives the lowest AIC value. See Sakamoto, Ishiguro and Kitagawa (1986) for more information about AIC and its applications.

3.5 Penalised Spline Smoothing

In penalised spline smoothing a relatively large number of knots are used, however, their influence is constrained by applying a penalty function. This approach is similar to that of *smoothing splines*, which date back to the work of Whittaker (1923). Smoothing splines minimise the penalised residual sum of squares

$$S = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_{x_{min}}^{x_{max}} f''(x)^2 dx. \tag{3.4}$$

The first term measures the closeness to the data while the second term penalises curvature in the function. The integrated squared second derivative in Equation

(3.4) has become popular as the choice of *roughness penalty* (Reinsch, 1967), although other penalties have been suggested (see Good and Gaskins, 1971; Boneva, Kendall and Stefanov, 1971). The parameter λ is the *smoothing parameter* controlling the amount of smoothing and establishes a trade-off between the closeness of the fit and the smoothness. Small values of λ produce wiggly curves while large values produce the smoothest functions that wiggle the least in response to fluctuations in the data. As λ tends to infinity the second derivative is constrained to zero and the least squares fit is obtained.

Smoothing splines have knots at each unique value of x, and control overfitting by using least-squares estimation with a roughness penalty. Therefore the dimension of the corresponding spline basis increases with sample size, making this approach to smoothing infeasible for very large data sets. This can be circumvented by reducing the spline basis to, for example, pseudo-splines (see Hastie, 1996; Wood, 2003). Alternatively, the idea of P-splines is to start with a rich but finite dimensional basis, but instead of parametric fitting a penalised (or ridge) regression is carried out.

3.5.1 B-splines

If the spline is built from truncated polynomials or basis splines (see de Boor, 1978) it has been shown that the actual number and location of knots has little influence on the performance of the fit (Ruppert, 2002). Basis splines, or B-splines for short, are a popular representation of curves first proposed by Schoenberg (1946). They are constructed from polynomial pieces joined at the knots. Once the knots are given, it is easy to compute the B-splines recursively, for any degree

of the polynomial (see de Boor, 1972). A B-spline basis of degree 3 is shown in

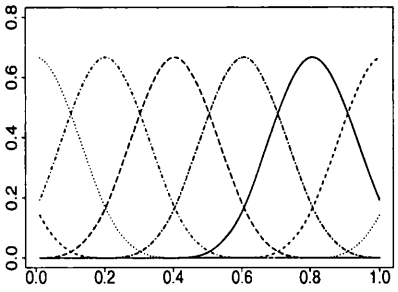


Figure 3.6: *A B-spline basis of degree 3.*

Figure 3.6.

Let $B_j(x)$ denote the value at x of the jth B-spline. When using the smoothness penalty suggested by Eilers and Marx (1996), the minimisation of S takes the following form

$$S = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{m} \eta_j B_j(x_i) \right\}^2 + \lambda \sum_{j=r+1}^{m} (\Delta^r \eta_j)^2.$$

The difference operator Δ^r is a discrete approximation to the integrated square of the rth derivative. This approach to smoothing reduces the dimensionality from n the number of observations, to m the number of B-splines, where m is chosen to be relatively large but much less than n.

3.5.2 Truncated Power Basis Function

The Eilers and Marx penalty can also be implemented using *truncated power functions* as the basis, provided that the penalty matrix is also transformed. The choice of penalty matrix will be discussed in the next Chapter. The *truncated*

power basis functions of degree d can be written as $1, x, \ldots, x^d, (x-t_1)_+^d, \ldots, (x-t_q)_+^d$. Figure 3.7 shows a truncated power basis of degree 3. The truncated power

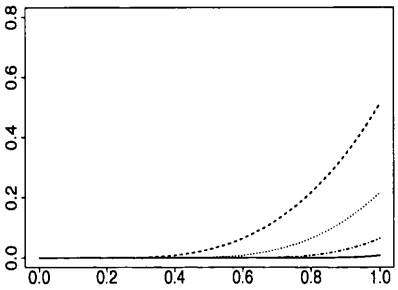


Figure 3.7: *A truncated power basis of degree 3.*

basis lacks the numerical stability of the B-spline basis and as a result the B-spline basis is often preferred in practice. However the truncated power basis can be used when the knots are carefully selected or when a penalised fit is applied (Ruppert et al., 2003).

3.5.3 Choice of Smoothing Parameter

One of the main objectives in penalised spline smoothing is choosing an appropriate smoothing parameter, λ. One approach is to choose λ subjectively, by comparing plots of the data with different smoothing parameters selected. Using this method one can vary the smoothing parameter and view features of the data in an exploratory fashion. Penalised spline fits of the data simulated in Equation (3.2) are shown in Figure 3.8. Third degree B-splines are used, with a second order penalty. In Figure 3.8(a) the smoothing parameter $\lambda = 0.001$, resulting in a fit which is fairly wiggly. As λ increases the fit becomes less wiggly and a smooth fit is obtained in Figure 3.8(c). However when λ is too large, see Figure 3.8(d), the resulting fit is over smoothed.

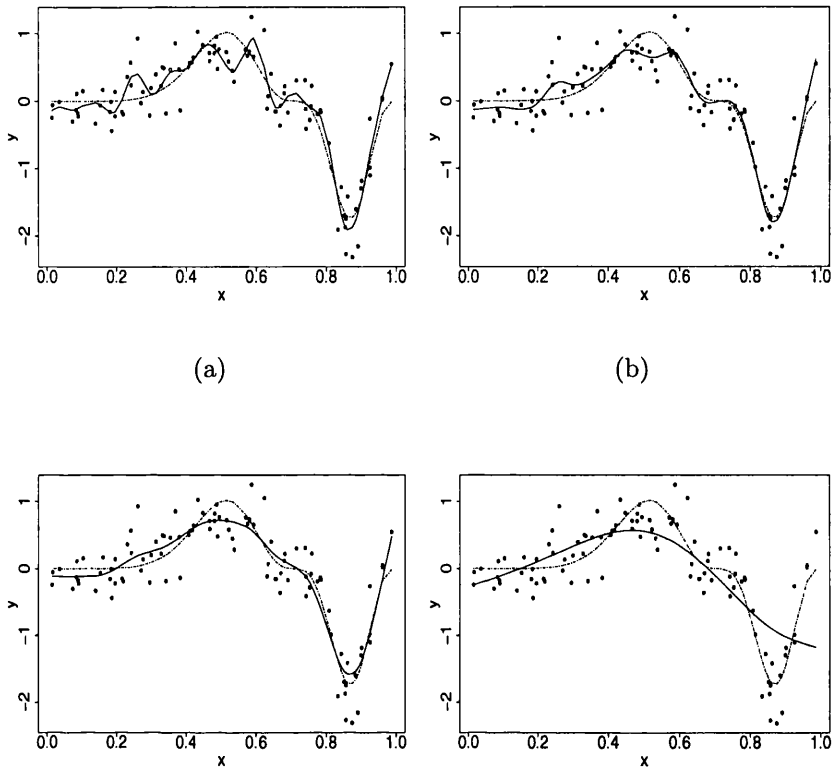


Figure 3.8: *A penalised spline fit with smoothing parameter (a)* $\lambda = 0.001$*; (b)* $\lambda = 0.1$*; (c)* $\lambda = 1$ *and (d)* $\lambda = 100$*. The true function is given by the dashed line.*

There is often a need for an automatic method whereby the smoothing parameter is chosen by the data. One such method used for the selection of λ is *cross-validation* (CV). Cross-validation predicts an observed value from remaining observations and chooses the λ that gives the best prediction. Recall the residual sum of squares in Equation (3.3) which is one of the most common measures for the goodness of fit. Let $\hat{f}(x;\lambda)$ denote the nonparametric regression estimate at point x with smoothing parameter λ. The residual sum of squares

can be written as

$$RSS(\lambda) = \sum_{i=1}^{n} \{y_i - \hat{f}(x_i; \lambda)\}^2.$$

Let $\hat{f}_{-i}$ denote the nonparametric regression estimator but with (x_i, y_i) removed from the data. Then the cross-validation criterion is

$$CV(\lambda) = \sum_{i=1}^{n} \{y_i - \hat{f}_{-i}(x_i; \lambda)\}^2.$$

One can show (see Efron, 1982; Ruppert et al., 2003) that

$$CV(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}_{-i}(x_i; \lambda)}{1 - S_{\lambda,ii}} \right)^2,$$

where $\mathbf{S}_\lambda$ is the smoother matrix associated with $\hat{f}$ and $S_{\lambda,ii}$ is its diagonal elements.

A simplified version known as *generalised cross-validation*, or GCV, was suggested by Craven and Wahba (1979):

$$GCV(\lambda) = \sum_{i=1}^{n} \left(\frac{\{(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}\}_i}{1 - n^{-1}\mathrm{tr}(\mathbf{S}_\lambda)} \right)^2,$$

This is a variant of the cross-validation approach in which the quantities $S_{\lambda,ii}$ are replaced by their average

$$\frac{1}{n} \sum_{i=1}^{n} S_{\lambda,ii} = \frac{1}{n} \mathrm{tr}(\mathbf{S}_\lambda).$$

The best λ is that which minimises CV(λ) or GCV(λ). Other smoothing parameter selection criteria exist such as Mallow's C_p criterion (Mallows, 1973) and Akaike's information criterion (AIC) (Akaike, 1973), however caution is needed

when using one of these methods. GCV and AIC selectors in particular lead to highly variable choices of smoothing parameters, and also possess a noticeable tendency towards under smoothing.

A relationship exists between P-splines and linear mixed models, and penalised quasi-likelihood (PQL) estimation in generalised linear mixed models (GLMMs) (see Breslow and Clayton, 1993; Schall, 1991) which may be utilised for smoothing parameter selection. For normal response models the link to GLMMs is discussed in the next Section.

3.6 Relationship between P-splines and Generalised Linear Mixed Models

For normal responses, Wand (2003) showed that linear mixed models can be used in the scatterplot smoothing setting. Consider a linear model of the form

$$y_i = a + b_1 x_i + \sum_{\kappa=1}^{q} \eta_\kappa (x_i - t_\kappa)_+ + \epsilon_i. \tag{3.5}$$

If the η_κ are treated as ordinary parameters, that is all coefficients are fixed effects, and estimated using ordinary least squares then the resulting fit will be quite rough due to the large number of knots being used. A solution is to treat the η_κ as independent *random effects*:

$$\eta_1, \ldots, \eta_q \sim N(0, \sigma_\eta^2). \tag{3.6}$$

For $\sigma_\eta^2 < \infty$ the η_κ are shrunk leading to a smooth fit and the number of basis functions is less than the number of observations, similar to penalised spline smoothing.

One can write the fixed effects vector as $\mathbf{b} = [a, b_1]^T$ and the random effects vector as $\eta = [\eta_1, \ldots, \eta_q]^T$. Now define the following matrices

$$\mathbf{X} = [1, x_i]_{1 \le i \le n} \qquad \text{and} \qquad \mathbf{Z} = [(x_i - t_\kappa)_+]_{1 \le i \le n,\, 1 \le \kappa \le q}.$$

Equations (3.5) and (3.6) can be rewritten as the linear mixed model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\eta + \epsilon, \quad \begin{bmatrix} \eta \\ \epsilon \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_\eta^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 \mathbf{I} \end{bmatrix} \right),$$

with $\mathbf{I}$ as identity matrix. The class of *generalised linear mixed models* is an extension of linear mixed models. This allows generalised responses from an exponential family, such as the Poisson or binomial distribution, to be considered. For *non-normal response models*, the link to GLMMs and corresponding smoothing parameter estimates will be discussed in greater detail in the next Chapter.

3.6.1 Penalised Quasi-Likelihood Estimation

Now consider P-spline smoothing in generalised models of the form

$$E(y|x) = g\{\nu(x)\}, \tag{3.7}$$

with $g(\cdot)$ as a known link function and $\nu(x)$ as a smooth but unknown function in x. The response y is assumed to follow an exponential family distribution with

$$f(y|\theta) \sim \exp\{y\theta - k(\theta)\},$$

where $\theta = \nu(x)$ is assumed to be a natural link. For P-spline fitting, replace $\nu(x)$ by $Xb + Z\eta$ where $X = (1, x)$ is a low dimensional basis in x and $Z = Z(x)$ is a high dimensional basis in x. The dimension of Z is large but fixed for increasing sample size. Coefficients η are estimated in penalised form by the penalised likelihood

$$l(\nu) - \frac{1}{2}\lambda\eta^T D\eta,$$

with $l(\nu) = \log f(y|\nu)$ as the likelihood and λ as the smoothing parameter steering the amount of penalisation. The matrix D is an appropriately chosen penalty matrix.

The estimates $\hat{b}$ and $\hat{\eta}$ are equivalent to penalised quasi-likelihood estimation in generalised linear mixed models as suggested in Breslow and Clayton (1993). Let η be independently normally distributed with

$$\eta \sim N(0, \sigma_\eta^2 D^-), \tag{3.8}$$

where D^- is the generalised inverse of D. The amount of smoothing is controlled by σ_η^2 and its reciprocal acts as the smoothing parameter λ. Conditional on η one models

$$y|\eta \sim f(y, \nu(\eta)) \tag{3.9}$$

with $\nu(\eta) = Xb + Z\eta$. Equations (3.8) and (3.9) provide the components of a

generalised linear mixed model. The likelihood for parameters η and λ are now obtained by integrating out the random effects η such that

$$f_y(y;\eta,\sigma_\eta^2) = \int f(y|\nu)\phi(\eta,\sigma_\eta^2 D^-)d\eta, \tag{3.10}$$

with $\phi(\cdot)$ as a normal density function. This integral is analytically intractable therefore an approximation is required. A convenient choice is a Laplace approximation of the integral leading to penalised quasi-likelihood estimates which are equivalent to penalised estimates in the smooth model (3.7).

Focus now lies on estimation of the a priori variance σ_η^2. Let

$$\hat{\eta} = \arg\max \log\{f(y|\nu(\eta))\phi(\eta,\sigma_\eta^2 D^-)\},$$

be the penalised estimate. The integral in Equation (3.10) can be approximated by

$$f_y(y;\eta,\sigma_\eta^2) \approx \frac{1}{\sqrt{|Z^T F Z - D/\sigma_\eta^2|}} f(y|\nu(\hat{\eta}))\phi(\hat{\eta},\sigma_\eta^2 D^-) \tag{3.11}$$

with $F = \partial^2 k(\theta)/\partial\theta$. Differentiating Equation (3.11) with respect to σ_η^2 provides the Laplace approximated maximum likelihood estimate through

$$\frac{\partial}{\partial\sigma_\eta^2}\log f_y(y;\eta,\sigma_\eta^2) = \frac{1}{2}\frac{\hat{\eta}^T D\hat{\eta}}{\sigma_\eta^4} - \frac{1}{2\sigma_\eta^2}df$$

with $df = \mathrm{tr}\{(Z^T F Z - D/\sigma_\eta^2)Z^T F Z\}$ as the degrees of freedom. This results in the Laplace based variance estimate

$$\hat{\sigma_\eta^2} = \frac{\hat{\eta}^T D\hat{\eta}}{df},$$

or in terms of the smoothing parameter, λ

$$\hat{\lambda} = \frac{df}{\hat{\eta}^T D \hat{\eta}}.$$

3.7 *P*-spline Smoothing in Survival Models

Allowing covariate effects to be dynamic across time leads to a varying-coefficient model as described by Hastie and Tibshirani (1993). The covariate effects are then smoothly estimated using, for example, local estimation as considered in Fan, Gijbels and King (1997) and Cai and Sun (2002) or spline fitting as considered by a number of authors including Zucker and Karr (1990), Kooperberg, Stone and Troung (1995), O'Sullivan (1988) and Strawderman and Tsiatis (1996). Gray (1992) used cubic splines to explore the functional form of the relationship between covariates and outcome with applications to breast cancer prognosis, and later, in testing hypothesis on covariate effects in a proportional hazards model (Gray, 1994). In that paper he also examines how covariate effects change over time. Rosenberg (1995) used *B*-splines to estimate the hazard function with application to acquired immunodeficiency syndrome (AIDS) following infection with human immunodeficiency virus (HIV). More recently *P*-splines have been used, in a mixed-model framework, by Cai, Hyndman and Wand (2002) for estimation of the hazard function with no covariates, an idea which is extended to proportional hazards models in Cai and Betensky (2003).

3.8 Chapter Summary

The concept of using a spline basis combined with a penalty is not new (O'Sullivan, 1986) but the procedure of penalised spline fitting has become extremely popular recently, mainly due to the paper by Eilers and Marx (1996). The underlying idea of P-spline smoothing is to fit a smooth curve by using a high dimensional basis which is then penalised to provide a smooth fit. The main difference between P-spline smoothing and smoothing splines is that with smoothing splines the dimension of the corresponding basis increases with sample size while in P-spline smoothing a rich but finite dimensional basis is used. P-spline smoothing also has strong similarities to linear mixed models and to penalised quasi-likelihood estimation in generalised linear mixed models. This link can be utilised for data driven selection of the smoothing parameter. In the next Chapter the first of two approaches to the smooth estimation of dynamic covariate effects is considered.

Chapter 4

A Poisson Based Approach to the Smooth Estimation of Dynamic Covariate Effects

4.1 Introduction

Estimation of covariate effects in the Cox proportional hazards model is usually based on the partial likelihood function. However, there are several reasons why in fact it may be worthwhile to work directly with the likelihood function when estimating dynamic covariate effects. In this Chapter the use of the likelihood function will be considered. A mixed-model approach is followed based on an extention of the work by Cai et al. (2002) and Cai and Betensky (2003) who use the mixed model idea for estimation of the hazard function with no covariates and for estimation in proportional hazards models respectively. Extending those ideas

enable estimates of dynamic covariate effects together with a smooth estimate of the baseline hazard to be obtained. Smooth estimates are obtained using P-spline smoothing.

4.2 The Likelihood Function

In the context of smooth estimation in survival models it is worth considering working directly with the likelihood function. Let T_i denote the survival time of the ith individual and let C_i be the corresponding right censored time, $i = 1, \ldots, N$. Observe $Y_i = \min(T_i, C_i)$ and define the censoring indicator as $\delta_i = 1$ if $T_i < C_i$ and $\delta_i = 0$ otherwise. When covariate effects are constant over time, the cumulative hazard function, $H(t)$, in the likelihood function factorises to the covariate effects multiplied by the cumulative baseline hazard as follows

$$\begin{aligned} H(t) &= \int_0^{Y_i} h_0(t) \exp(\beta^T x_i) dt \\ &= \exp(\beta^T x_i) \int_0^{Y_i} h_0(t) dt. \end{aligned}$$

The baseline hazard is then be estimated using one of the methods described in Section 2.5 and the resulting profile likelihood for the parameters is equivalent to the partial likelihood (Cox, 1972). If however, covariate effects do vary with time then such factorisation of the cumulative hazard does not exist and the partial likelihood does not have any justification as the profile likelihood function. Hence use of the likelihood function should be considered in this case. Another reason to work with the likelihood function is that one may obtain a smooth estimate of the baseline hazard function, usually treated as a nuisance parameter. The

dynamic hazard function

$$h(t|\beta, x_i) = h_0(t) \exp(\beta^T(t) x_i),$$

can be easily rewritten in order to incorporate an estimate of the baseline hazard function when dynamic covariate effects $\beta(t)$ are estimated.

4.3 Penalised Likelihood Function

Rewrite the dynamic hazard function as

$$h(t|\beta, x_i) = \exp(\beta^T(t) z_i), \tag{4.1}$$

with $z_i^T = (1, x_i^T)$, and $\beta(t) = \{\beta_0(t), \beta^T(t)\}^T\}$ where $\beta_0(t) = \log h_0(t)$ is the baseline hazard function. These smooth estimates are obtained using P-spline smoothing. Define a high dimensional spline basis $B(t) = \{b_1(t), \ldots, b_q(t)\}$ over knots $t_1, \ldots, t_q$. A large number of knots are chosen so that the fit shows more variation than is justified by the data however q will typically be far less then the number of observations N. Ruppert (2002) showed that the actual number of knots chosen have little influence on the fit however it is generally believed that too many knots are better then too few knots. This is because once enough knots have been selected to fit the data, overfitting is controlled by the penalty function. Knots may be placed evenly such that the distance between any two adjacent knots is the same (Eilers and Marx, 1996) or, alternatively, placed at quantiles of x so that there are about the same number of observed values between any two adjacent knots (Ruppert et al., 2003). Equally spaced knots are used

here (see Eilers and Marx (2004) who show that regardless of whether B-splines or truncated splines are used there is no need for unequally spaced knots).

Consider the smooth estimation of the baseline function $\beta_0(t)$. The penalty function used for this component is

$$\lambda_0 \alpha_0^T D_0 \alpha_0,$$

where λ_0 is the smoothing parameter controlling the amount of smoothing, D_0 is an appropriately chosen penalty matrix and α_0 are the basis coefficients.

4.3.1 Penalty Matrix

The penalty matrix D chosen depends on the basis functions used. Eilers and Marx (1996) used difference based penalties with B-spline basis functions. This form of penalty is based on finite differences of the coefficients of adjacent B-splines. First order differences $\Delta\alpha_j$ of the jth B-spline are written as

$$\Delta\alpha_j = \alpha_j - \alpha_{j-1},$$

or as a matrix $\Delta\alpha = D\alpha$, where

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

Note that it is possible to use different orders of the differences. When truncated polynomials are used, a reasonable choice of penalty matrix is the identity matrix (see Wand, 2003). The identity matrix is defined as a diagonal matrix with 1 in every entry of its main diagonal. Therefore penalty matrix D is simply

$$D = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

4.3.2 Smooth Estimation

The aim is to smoothly estimate $\beta_l(t)$, $l = 0, \dots, p$. It is useful to extract the intercept from the smooth function so that $\beta_l(t) = \beta_{0l} + \tilde{B}(t)\alpha_l$. Here β_{0l} is the constant fit while $\tilde{B}(t)$ is the basis matrix containing no intercept. Define $\theta_l = (\beta_{0l}, \alpha_l^T)^T$ as the parameter vector estimating the linear fit and the basis coefficients. A starting value for the penalised fit can be obtained by fitting a model with an unpenalised baseline hazard and with covariate effect estimates based on a Cox-PH fit. Using the The Kronecker product (an operation on two matrices of arbitrary size resulting in a larger block matrix), denoted $\otimes$, one can write $\beta(t) = \mathbf{W}(t)\boldsymbol{\theta}$ where $\mathbf{W}(t) = I_{p+1} \otimes \{1, \tilde{B}(t)\}$ (I_{p+1} is the $p+1$ dimensional identity matrix) and parameter vector $\boldsymbol{\theta} = (\theta_0^T, \dots, \theta_p^T)$. The spline bases used in each component of $\beta(t)$ may differ so that $\mathbf{W}(t)$ is of block diagonal form with different spline bases on its diagonal. However for simplicity of notation let $\mathbf{W}(t)$ be constructed using the same spline bases functions.

One can write the likelihood function in the smooth context as

$$l_i(\boldsymbol{\theta}) = \delta_i \left(z_i^T \mathbf{W}(Y_i)\boldsymbol{\theta} \right) - \int_0^{Y_i} \exp \left\{ z_i^T \mathbf{W}(t)\boldsymbol{\theta} \right\} dt,$$

for $i = 1, \ldots, N$. The coefficients α_l are then jointly penalised leading to the *penalised log likelihood function*

$$l^P(\boldsymbol{\theta}, \lambda) = \sum_{i=1}^{N} l_i(\boldsymbol{\theta}) - \frac{1}{2} \sum_{l=0}^{p} \lambda_l \alpha_l^T D_l \alpha_l. \tag{4.2}$$

Each component has its own penalty function. Component-wise smoothing parameters are $\lambda_l = (\lambda_0, \ldots, \lambda_p)$. The penalty in Equation (4.2) can be written as $\boldsymbol{\theta}^T(\boldsymbol{\Lambda}\mathbf{D})\boldsymbol{\theta}$ for notational convenience. Here $\mathbf{D}$ is a block diagonal matrix built from matrices $\text{diag}(0, D_l)$, $l = 0, \ldots, p$. Let 1_q be the q dimensional unit vector. The smoothing matrix $\boldsymbol{\Lambda}$ is also a diagonal matrix with $(\lambda_0 \otimes 1_q^T, \ldots, \lambda_p \otimes 1_q^T)$ as diagonal elements. Differentiating the penalised likelihood function with respect to $\boldsymbol{\theta}$ gives the penalised score function

$$\frac{\partial l^P(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\mathbf{D}\theta, \tag{4.3}$$

with $s_i(\boldsymbol{\theta}) = \delta_i(\mathbf{W}^T(Y_i)z_i) - \int_0^{Y_i} \mathbf{W}^T(t)z_i \exp\{z_i^T \mathbf{W}(t)\boldsymbol{\theta}\}dt$. On differentiating the penalised score equation one obtains the second order derivative

$$\frac{\partial^2 l^P(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\mathbf{D}, \tag{4.4}$$

where $\nabla s_i(\boldsymbol{\theta}) = -\int_0^{Y_i} \mathbf{W}^T(t) z z_i^T \mathbf{W}(t) \exp\{z_i^T \mathbf{W}(t)\boldsymbol{\theta}\}dt$.

4.4 Numerical Integration

The penalised score function in Equation (4.3) and the second order derivative in Equation (4.4) contain integrals based on the hazard function. Therefore the integrals have to be approximated using numerical integration. A number of methods are available but a computationally handy method to use is *trapezoid integration.*

Trapezoid integration approximates a definite integral $\int_a^b f(x)dx$. The interval $a \leq b$ is divided into n equal subintervals. Consider the function $f(x) = 2x_i^2 sin^3(2\pi x_i^2)$ divided into subintervals each of width w, as shown in Figure 4.1. The area of each subinterval may be obtained approximately by assuming that

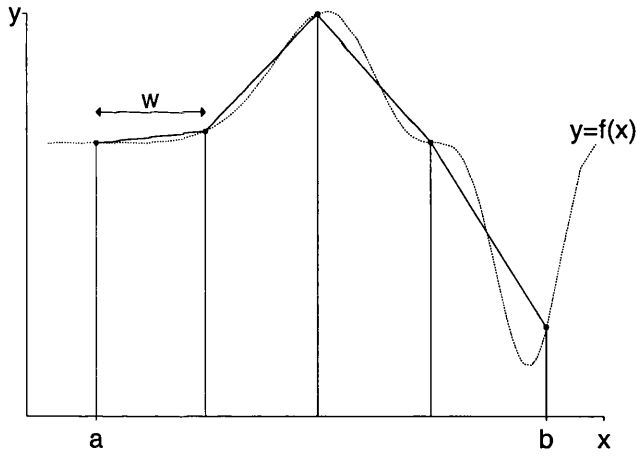


Figure 4.1: *An illustration of trapezoid integration. The function* $f(x) = 2x_i^2 sin^3(2\pi x_i^2)$ *is divided into equal intervals of width* w. *The area of each subinterval is assumed to be the union of a rectangle and a triangle.*

the subinterval is a trapezoid, which is simply the union of a rectangle and a triangle. The trapezoid rule approximation to the integral over the entire interval

is

$$\int_a^b f(x)dx \approx \frac{1}{2}w\{f(x_0) + 2f(x_1) + \ldots + 2f(x_{n-1}) + f(x_n)\}.$$

Now define $0 = \tau_0 < \tau_1 < \ldots < \tau_K$ to be a grid of points spanning the range of observed failure times where $\tau_1 = \min\{Y_i : \delta_i = 1\}$, and $\tau_K = \max\{Y_i : \delta_i = 1\}$. Define K_i such that $\tau_{K_i-1} < Y_i \leq \tau_{K_i}$. Let $\mathbf{m}_i(t) = \mathbf{W}^T(t)z_i \exp\{z_i^T\mathbf{W}(t)\boldsymbol{\theta}\}$ be the integrand in Equation (4.3) which can be approximated using trapezoid integration.

$$\begin{aligned}
&\int_0^{Y_i} \mathbf{m}_i(t)dt \\
&\approx\ d(K_i > 1) \sum_{k=1}^{K_i-1} \frac{1}{2}(\tau_k - \tau_{k-1})\{\mathbf{m}_i(\tau_k) + \mathbf{m}(\tau_{k-1})\} \\
&\quad + \frac{1}{2}(Y_i - \tau_{K_i-1})\{\mathbf{m}_i(\tau_{K_i-1}) + \mathbf{m}(\tau_{K_i})\} \\
&= \frac{1}{2}\min(\tau_1, Y_i)\mathbf{m}_i(\tau_0) + \frac{1}{2}\sum_{k=1}^{K_i}\{\min(\tau_{k+1}, Y_i) - \min(\tau_{k-1}, Y_i)\}\mathbf{m}_i^T(\tau_k),
\end{aligned}$$

where $d(\cdot)$ is an indicator function. Using trapezoid integration, the score function $\mathbf{s}_i(\boldsymbol{\theta})$ can now be approximated by

$$\mathbf{s}_i(\boldsymbol{\theta}) = \delta_i(\mathbf{W}^T(Y_i)z_i) - \sum_{k=0}^{K_i} \mathbf{W}^T(\tau_k)z_i \exp\{z_i^T\mathbf{W}(\tau_k)\boldsymbol{\theta} + o_{ik}\}, \qquad (4.5)$$

with the coefficient of the term o_{ik} equal to 1. In the same way, the unpenalised second order derivative $\nabla\mathbf{s}_i(\boldsymbol{\theta})$ is approximated by

$$\nabla\mathbf{s}_i(\boldsymbol{\theta}) = -\sum_{k=0}^{K_i} \mathbf{W}^T(\tau_k)z_i z_i^T\mathbf{W}(\tau_k) \exp\{z_i^T\mathbf{W}(\tau_k)\boldsymbol{\theta} + o_{ik}\}.$$

Based on standard expansions one can take advantage of a *sandwich type*

estimator for variance estimation of parameter estimates

$$\text{var}(\hat{\boldsymbol{\theta}}) = -\left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\mathbf{D}\right\}^{-1} \left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta})\right\} \left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\mathbf{D}\right\}^{-1}.$$

The virtue of the sandwich covariance estimator is that it provides consistent estimates of the covariance matrix for parameter estimates even when the fitted model fails to hold or is left unspecified. For the lth fitted component, $l = 0, \ldots, p$, the variance estimator for the functional shape is then obtained as $\text{Var}(\hat{\beta}_l(t)) = \text{Var}(W_l(t)\hat{\boldsymbol{\theta}}_l)$, which is equivalent to $\text{Var}(\hat{\beta}_l(t)) = W_l(t)\text{Var}(\hat{\theta}_l)W_l^T(t)$, and a pointwise 95% confidence interval for the lth component is given as

$$\text{C.I.}(\hat{\beta}_l(t)) = \hat{\beta}_l(t) \pm 1.96\sqrt{\text{Var}(\hat{\beta}_l(t))}.$$

4.4.1 Poisson Regression Model

Cai et al. (2002) showed that a penalised spline hazard estimate can be approximated by a Poisson model with an offset. In the same way here, approximation (4.5) shows the form of a Poisson model fitted to independent pseudo observations $Y_{ik} \sim Po(z_i\mathbf{W}(\tau_k)\boldsymbol{\theta} + o_{ik})$, $k = 1, \ldots, K_i$, $i = 1, \ldots, N$, with $Y_{ik} = 0$ for $k < K_i$ and $Y_{iK_i} = \delta_i$. The corresponding offset values are as follows. For $k = 0$, one can define $o_{i0} = \log\{\min(\tau_1, Y_i)\}$ and for $k = 1, \ldots, K_i$ $o_{ik} = \log[\frac{1}{2}\{\min(\tau_{k+1}, Y_i) - \min(\tau_{k-1}, Y_i)\}]$. Hence using numerical integration the dynamic hazard function in Equation (4.1) can be approximated by fitting a penalised Poisson regression model with given offset o_{ik} and pseudo data Y_{ik}.

4.4.2 Grid Point Selection

In practice, the number of integration grid points as well as their location must be chosen. A coarser grid omits information in the data while a finer grid will leads to identifiability problems. In choosing their location, it makes sense to use the observed failure times with grid point τ_0 equal to 0, grid point τ_1 equal to the first event time, τ_K equal to the last event time and τ_{K+1} set to infinity. All other integration grid points are equally spaced between τ_1 and τ_K. The number of grid points chosen should be at least as large as the number of knots q in order to achieve identifiability.

4.5 Link to Generalised Linear Mixed Models

4.5.1 Penalised Quasi-Likelihood Estimation

In the previous Chapter the relationship between P-spline smoothing and penalised quasi likelihood (PQL) estimation in generalised linear mixed models (GLMMs) was discussed. In particular the link for normal response models (Wand, 2003) was illustrated. For non-normal response models the link is achieved in the following way. Consider coefficients α_l as independent normally distributed variables with

$$\alpha_l \sim N(0, \lambda_l^{-1} D_l^-), \tag{4.6}$$

where D_l^- is the generalised inverse of D_l. Smoothing parameters λ_l now occur in the *a priori* variance of α_l. Conditional on α_l, $l = 0, \ldots, p$, and based on the

trapezoid integration the response Y_{ik} is modelled as

$$Y_{ik}|(\alpha_l) \sim Po(Y_{ik}; z_i^T \mathbf{W}(\tau_k)\boldsymbol{\theta} + o_{ik}), \tag{4.7}$$

Equations (4.6) and (4.7) provide the components of a generalised linear mixed model with random effects α_l. To obtain the likelihood for β_{0l} and smoothing parameters λ_l, $l = 0, \ldots, p$, one needs to integrate out the random coefficients α_l as follows

$$l(\beta_{0l}, \lambda_l) = \int \prod_{i=1}^{N} \prod_{k=0}^{K_i} Po(Y_{ik}; z_i^T \mathbf{W}(\tau_k)\boldsymbol{\theta} + o_{ik}) \times \prod_{l=0}^{p} \phi(\alpha_l, \lambda_l^{-1} D_l^{-1}) d\alpha_l, \tag{4.8}$$

with $\phi(\cdot)$ as a normal density function. Using a Laplace approximation for the integral in Equation (4.8) leads to penalised quasi-likelihood estimation (Breslow and Clayton, 1993). That is to say, that Laplace approximation based estimates for β_{0l} (and α_l) in the generalised linear mixed model are equivalent to penalised estimates in the smooth model (4.3).

The connection between P-spline smoothing and PQL estimation in GLMMs can be used for estimating the smoothing parameters λ_l, $l = 0, \ldots, p$. The estimation of λ_l is based on the likelihood function (4.8) and is obtained by approximating the integral using Laplace integration. Inserting estimates for β_{0l} gives the Laplace approximation for the log profile likelihood as

$$\begin{aligned} l^p(\lambda_l) &= \sum_{i=1}^{N} \sum_{k=0}^{K_i} \log Po(Y_{ik}; z_i^T \mathbf{W}(\tau_k)\boldsymbol{\theta} + o_{ik}) - \frac{1}{2} \sum_{l=0}^{p} (\lambda_l \hat{\alpha}_l^T D_l \alpha_l + \log|\lambda_l D_l|) \\ &\quad -\frac{1}{2} \log|\sum_{i=1}^{N} \mathbf{U}_i^T V_i \mathbf{U}_i + \mathrm{diag}(\lambda_l D_l)|, \end{aligned}$$

with $V_i = \mathrm{diag}(\mathrm{var}(Y_{i0}), \ldots, \mathrm{var}(Y_{iK_i}))$ and $\mathbf{U}_i$ as the observed design matrix of

pseudo Poisson variables for the ith individual. Maximising this with respect to the smoothing parameters yields the following PQL estimate of λ_l

$$\hat{\lambda}_l = \frac{\mathrm{df}_l}{\hat{\alpha}_l^T D_l \hat{\alpha}_l}, \tag{4.9}$$

where df_l is the approximate degrees of freedom for the lth smooth component given by

$$\mathrm{df}_l = \mathrm{tr}\left\{ (\sum_{i=1}^{N} z_{il}^2 B_i^T V_i B_i + \lambda_l D_l)^{-1} \sum_{i=1}^{N} z_{il}^2 B_i^T V_i B_i \right\},$$

and $B_i = (B^T(\tau_0), \ldots, B^T(\tau_{K_i}))^T$. The smoothing parameter estimate (4.9) depends on estimates $\hat{\boldsymbol{\theta}}$ and vice versa. One way to estimate both λ and $\boldsymbol{\theta}$ is to cycle between estimation of $\boldsymbol{\theta}$ for given λ and estimation of λ for given $\boldsymbol{\theta}$. Denote the jth cycle of such an algorithm as $\hat{\lambda}^{(j)}$ and $\hat{\boldsymbol{\theta}}^{(j)}$.

4.5.2 Hybrid Smoothing Parameter Selection

Breslow and Lin (1995) and Shun and McCullagh (1995) showed that Laplace approximation or PQL estimation of the marginal likelihood can perform poorly. In addition to the PQL estimate, which tends to over smooth the data, Kauermann (2003) considered the Monte Carlo EM algorithm as suggested by Booth and Hobert (1999). The algorithm works well in this setting however the improvement in fit is bought at the price of increased numerical effort. Finally a hybrid strategy based on an AIC choice was considered. This hybrid method has the advantage that it has the numerical simplicity of the PQL estimate, but with the possibility of over smoothing being controlled using the Akaike criterion. One should keep in mind that the smoothing parameters could in principle be data driven based on any common criterion such as the Akaike criterion, for exam-

ple. In practice this would be pursued by grid searching leading to a formidable exercise if p, the dimension of the smoothing parameters, is large. Therefore λ is estimated in a hybrid way by taking advantage of the numerical simplicity of the PQL estimate given in (4.9) but controlling smoothing parameters using the Akaike information criterion. It works by calculating the Akaike criterion, $AIC(\hat{\lambda}^{(j)})$, at the jth cycle of the PQL estimation, with

$$AIC(\lambda) = -2\sum_{i=1}^{N}\sum_{k=0}^{K_i} \log Po(Y_{ik}; z_i^T \mathbf{W}(\tau_k)\hat{\boldsymbol{\theta}} + o_{ik}) + 2\mathrm{df}(\lambda),$$

where $\mathrm{df} = \sum_{l=0}^{p} \mathrm{df}_l$ is the degree of freedom of the model. The iterations are ended if $AIC(\lambda^{j+1}) > AIC(\lambda^{(j)})$. In this way, an estimated smoothing parameter $\hat{\lambda} = (\hat{\lambda}_0, \ldots, \hat{\lambda}_p)$ is obtained which provides a small (though not minimal) value of the Akaike information criterion. At the final estimate $\hat{\lambda}^{(j)}$ a local grid search could now be run to find the minimum of $\mathrm{AIC}(\lambda)$. For ease of numerical effort, however, this step is omitted. Further ideas for multiple smoothing parameter selection are given in Wood (2000).

4.6 Simulations

4.6.1 Single Covariate

This Chapter considers the use of a Poisson type approach to estimate smooth dynamic covariate effects. Smooth estimates are achieved using P-spline smoothing. The behaviour of the approach will now be demonstrated using simulations. First consider simulating survival data for $N = 400$ individuals on a discrete time

grid, $t = 1, 2, \ldots, 60$. In each simulation, a single binary covariate is randomly chosen such that $P(x_1 = 1) = 0.5$. Let the effect of the covariate be either constant with $\beta(t) = 1$, or be one of the dynamic effects $\beta(t)$ described in Table 4.1 and shown in Figure 4.2. The baseline hazard function is constant, $\beta_0(t) = -5$.

Table 4.1: *Simulated Covariate Effects*

Dynamic Effects	$\beta(t)$
Linear (steep)	$\beta(t) = 2 * (1 - (t/50))$
Linear (flat)	$\beta(t) = t/50$
Cosinus	$\beta(t) = 0.75 * \cos(3t/40)$
Quadratic	$\beta(t) = -0.5 + 4t/100 - 4t^2/10000$

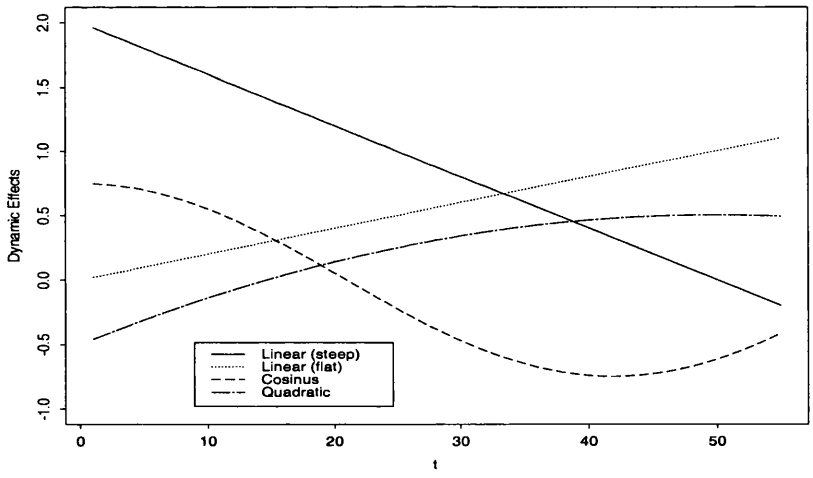


Figure 4.2: *Dynamic Effects corresponding to Table 4.1.*

Censoring is simulated to be independent of the covariate with a drop out probability of 3% or 0.1% at each time interval t to $t+1$. This means that, ignoring the event probability, around 16% and 95%, respectively, of the simulated individuals are expected to be in the simulated study until the end. In the latter case, the censoring rate is small but is included to demonstrate the effect of censoring in

general. For a constant covariate effect $\beta(t) = 1$, Figure 4.3 shows the estimated survivor function (left plots) and separate Kaplan-Meier estimates of the survivor function (right plot) for two groups where group 0 corresponds to the baseline group. The top row corresponds to a drop out probability of 3% at each time interval t to $t+1$, and the bottom row to a drop out probability of 0.1% at each time interval t to $t+1$. Estimates are based on a fitted Cox proportional hazards model.

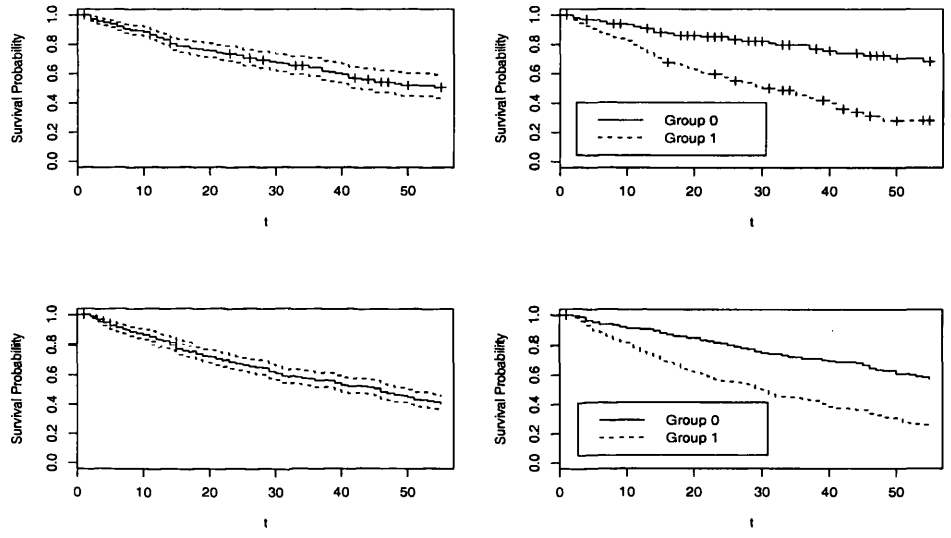
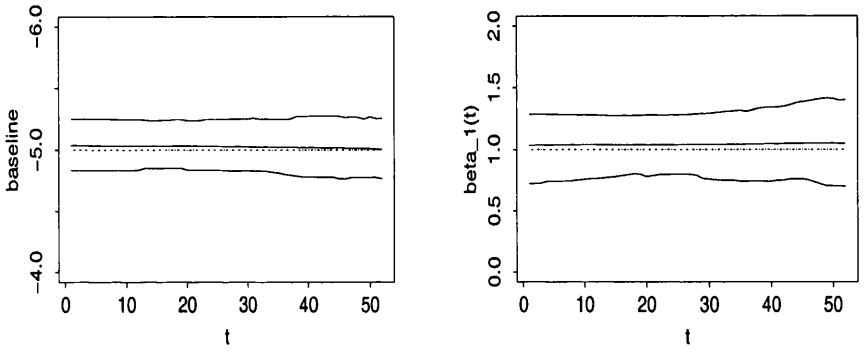


Figure 4.3: *Kaplan-Meier estimates (left plots) and Kaplan-Meier estimates by status (right plots) corresponding to a sample size of* $N = 400$. *In the top row, the drop out probability is* 3% *at each time interval* t *to* $t+1$, *in the bottom row, the drop out probability is* 0.1% *at each time interval* t *to* $t+1$.

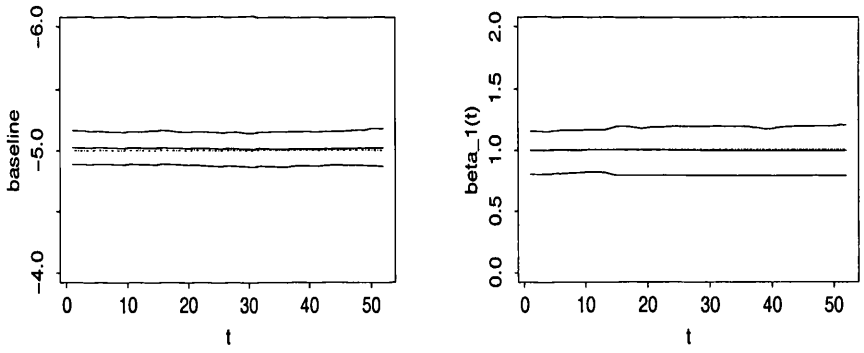
In determining the degree of spline to use, linear, quadratic and cubic truncated splines were compared with no real difference found between approaches. Therefore for simplicity, linear truncated splines are used here. In general, an increase in the degree of spline used will tend to lead to an increased fit, particularly in fitting smooth corners, however, if one uses enough knots then the

difference between, for example, a linear and quadratic spline fit is usually negligible. The number of knots q chosen in line with the guidelines given by Ruppert et al. (2003) with the number of grid points chosen such that $K = q$. Smoothing parameter selection is started with $\hat{\lambda}_l^{(0)} = 100$ for $l = 0, 1$ and is updated with $\lambda_l^{(t)}$ as long as the Akaike criterion decreases. In Figure 4.4 one can see the simulations for a time constant covariate effect, $\beta(t) = 1$, with a constant baseline hazard, $\beta_0(t) = -5$. Shown are the mean estimates from 150 simulations with 0.05 and 0.95 quantiles forming the bounds of a 90% confidence interval. Simulations for each of the dynamic effects are shown in Figures 4.5 to 4.8 respectively.

Simulated coverage probabilities are given in Figures 4.9 to 4.13. These are shown for the effects of covariates (baseline simulated coverage probabilities are similar to that obtained for a constant covariate effect). Plots for the dynamic effects given in Figures 4.10 to 4.13 show that simulated coverage probabilities are not as high as those for the constant effect. For both the steep linear and quadratic covariate effects (Figures 4.10 and 4.13), simulated coverage probabilities are low to begin with but increase with time. From Figures 4.5 and 4.8, there appears to be some differences between the estimated mean effect and the true underlying effect from time point 0 until around time point 10. Since there are many events around this time variance bands tend to be narrow, hence this could explain the low simulated coverage probabilities visualised early on for these effects. Hence a time constant effect is captured best. A lower drop out probability at each time interval t to $t + 1$ makes little difference to the simulated coverage probabilities obtained, however increasing sample sizes may help to capture dynamic effects more accurately.

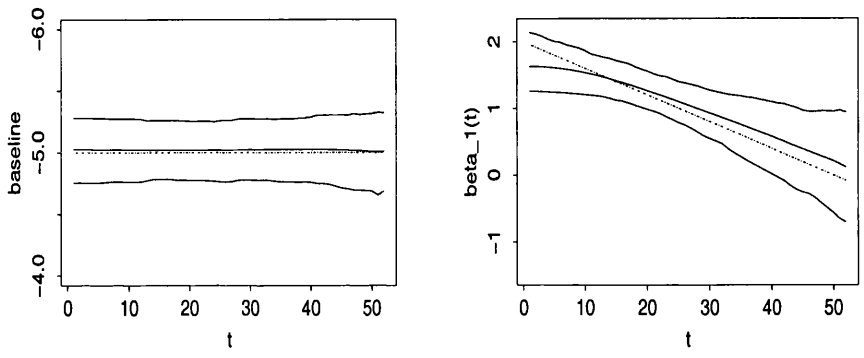


(a)

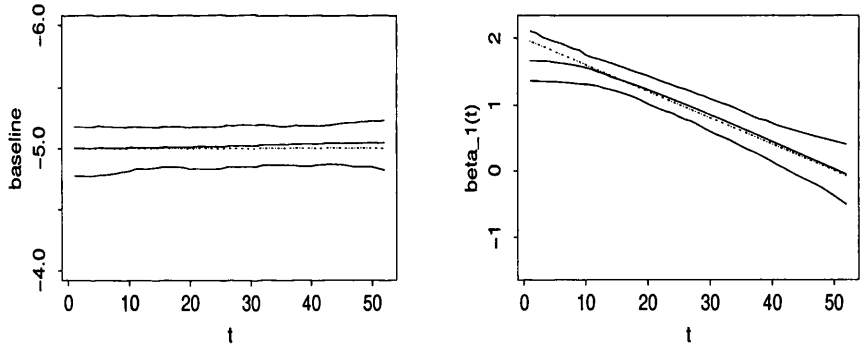


(b)

Figure 4.4: *P-spline Poisson type estimates of a constant covariate effect* ($\beta(t) = 1$)*, with a constant baseline hazard* ($\beta_0(t) = -5$)*. Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row (a) corresponds to a drop out probability of* 3% *at each time interval* t *to* $t + 1$*. The bottom row (b) corresponds to a drop out probability of* 0.1% *at each time interval* t *to* $t + 1$*. True functions are given by the dashed lines.*

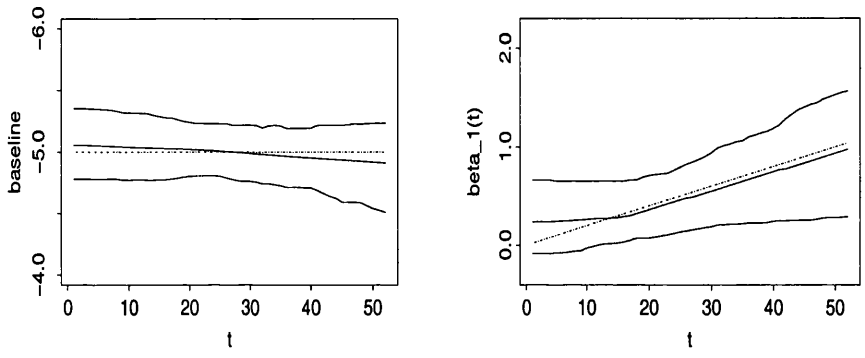


(a)

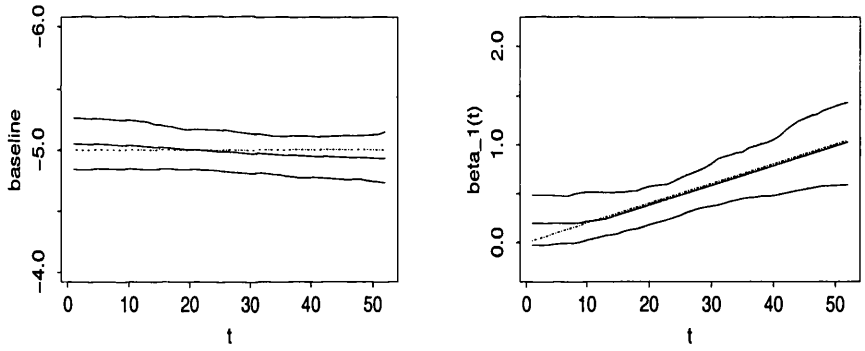


(b)

Figure 4.5: *P-spline Poisson type estimates of a linear (steep) covariate effect, with a constant baseline hazard ($\beta_0(t) = -5$). Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row (a) corresponds to a drop out probability of* 3% *at each time interval t to $t+1$. The bottom row (b) corresponds to a drop out probability of* 0.1% *at each time interval t to $t+1$. True functions are given by the dashed lines.*

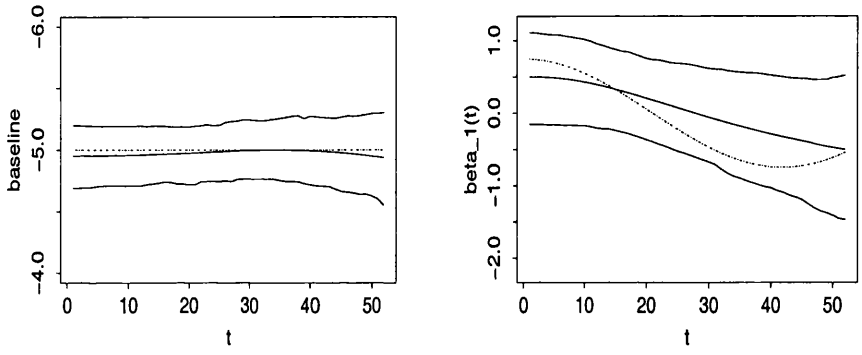


(a)

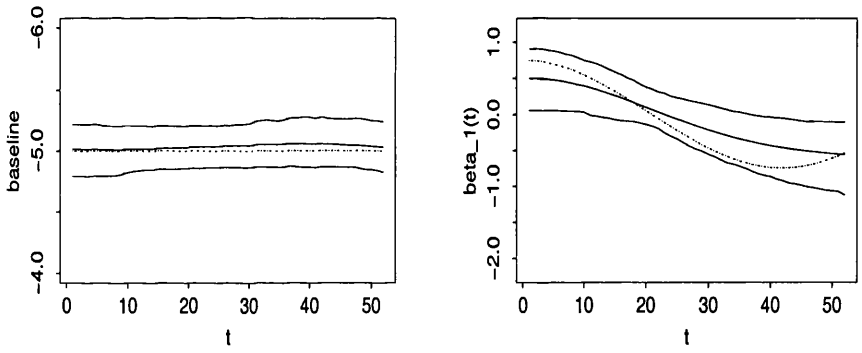


(b)

Figure 4.6: *P-spline Poisson type estimates of a linear covariate effect, with a constant baseline hazard* ($\beta_0(t) = -5$). *Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row (a) corresponds to a drop out probability of* 3% *at each time interval* t *to* $t+1$. *The bottom row (b) corresponds to a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *True functions are given by the dashed lines.*

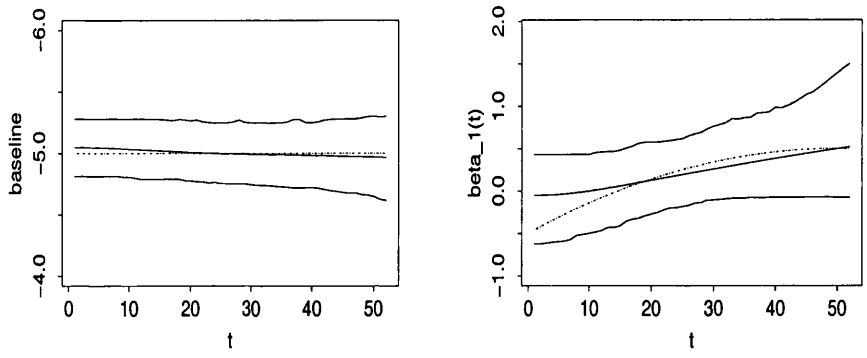


(a)

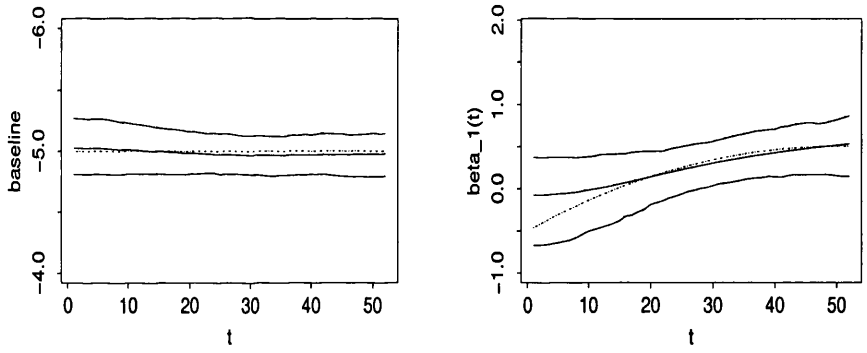


(b)

Figure 4.7: *P-spline Poisson type estimates of a cosinus covariate effect, with a constant baseline hazard* ($\beta_0(t) = -5$). *Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row (a) corresponds to a drop out probability of* 3% *at each time interval* t *to* $t + 1$. *The bottom row (b) corresponds to a drop out probability of* 0.1% *at each time interval* t *to* $t + 1$. *True functions are given by the dashed lines.*



(a)



(b)

Figure 4.8: *P-spline Poisson type estimates of a quadratic covariate effect, with a constant baseline hazard ($\beta_0(t) = -5$). Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row (a) corresponds to a drop out probability of* 3% *at each time interval t to $t+1$. The bottom row (b) corresponds to a drop out probability of* 0.1% *at each time interval t to $t+1$. True functions are given by the dashed lines.*

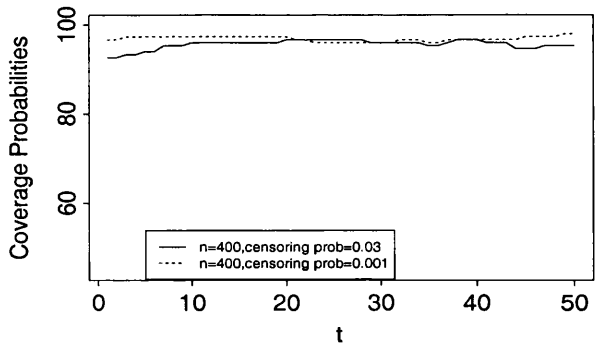


Figure 4.9: *Plot of coverage probabilities when the covariate effect is constant. Nominal value is* 95%.

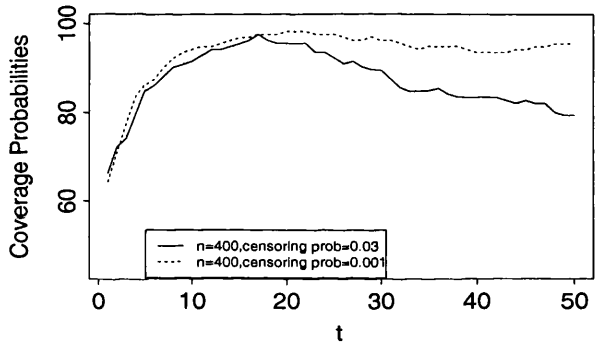


Figure 4.10: *Plot of coverage probabilities when the covariate effect is linear (steep). Nominal value is* 95%.

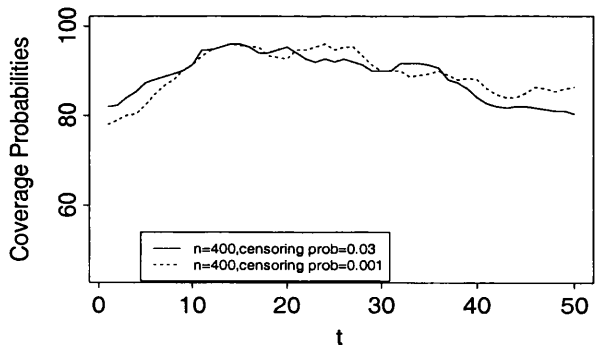


Figure 4.11: *Plot of coverage probabilities when the covariate effect is linear. Nominal value is* 95%.

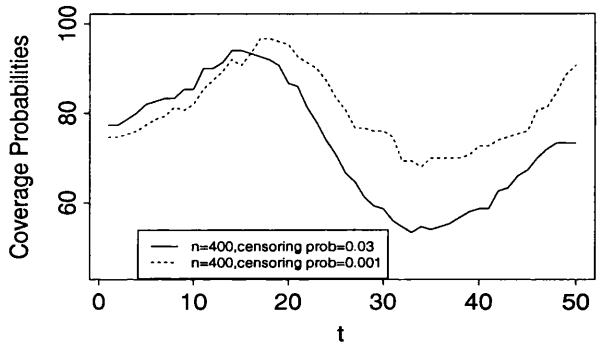


Figure 4.12: *Plot of coverage probabilities when the covariate effect is cosinus. Nominal value is* 95%.

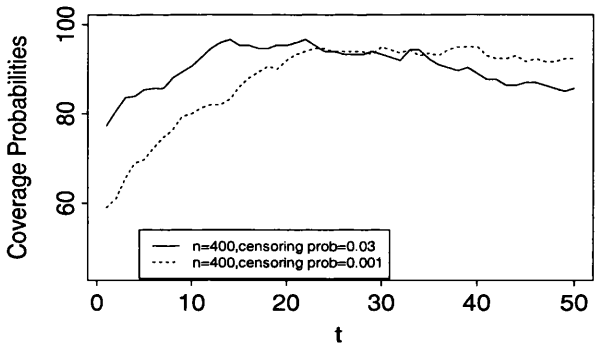


Figure 4.13: *Plot of coverage probabilities when the covariate effect is quadratic. Nominal value is* 95%.

4.6.2 Multiple Covariates

Now consider survival data for $N = 400$ individuals, simulated with two covariates on a discrete time grid, $t = 1, 2, \ldots, 60$. Both covariates are randomly chosen with $P(x_1 = 1) = 0.5$ and $P(x_2 = 1) = 0.3$. For ease of presentation consider only data simulated with a drop out probability of 3% at each time interval t to $t+1$. The covariate effects to be simulated, in each of three settings, are shown in Table 4.2. The baseline hazard function is included as a constant effect such that $\beta_0(t) = -5$. Smoothing parameter selection is started with $\hat{\lambda}_l^{(0)} = 100$ for $l = 0, 1, 2$ and is updated with $\lambda_l^{(t)}$ as long as the Akaike criterion decreases.

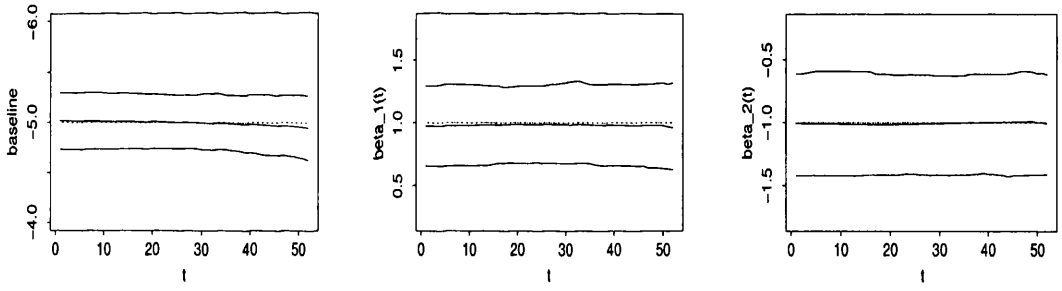
Table 4.2: *A 2 Covariate Scenario*

	Covariate 1	Covariate 2
1	$\beta_1(t) = 1$	$\beta_2(t) = -1$
2	$\beta_1(t) =$ cosinus effect	$\beta_2(t) = 1$
3	$\beta_1(t) =$ quadratic effect	$\beta_2(t) =$ cosinus effect

The means of 150 simulations, with corresponding pointwise 90% confidence intervals, are shown in Figures 4.14 to 4.16. Figure 4.14 gives the mean estimates and pointwise 90% confidence intervals for survival data simulated with two constant covariates. Figure 4.15 gives corresponding output for survival data simulated with one dynamic covariate effect and one time constant covariate effect and Figure 4.15 corresponding output for survival data simulated with two dynamic covariate effects. In each case the baseline hazard function is kept constant.

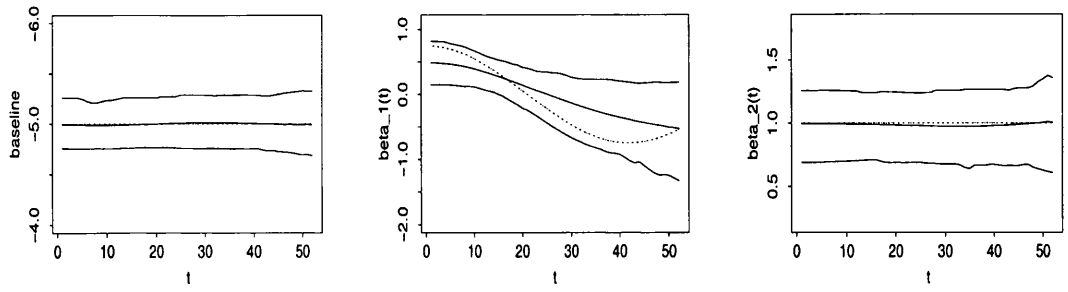
In Figure 4.17 the steps taken by the smoothing parameter, for one simulation,

are shown. The smoothing parameter selection procedure is started each time with $\hat{\lambda}_l^{(0)} = 100$, $l = 0, 1, 2$. Final estimates are based on the Akaike criterion with steps terminated when $AIC(\lambda^{(j+1)}) > AIC(\lambda^{(j)})$. Around five to six steps are generally needed to achieve the data driven smoothing parameter estimate. The plots in Figure 4.17 show the steps taken by λ_1 and λ_2 (with λ_0 reaching infinity due to the constant baseline hazard). In Figure 4.17(a) the steps for two constant covariates are shown. Here, $\hat{\lambda}_l \to \infty$, $l = 1, 2$, meaning that two proportional hazard fits are obtained. A smoothing parameter estimate which does not go to infinity indicates that a dynamic covariate effect has been captured. This is visible in Figure 4.17(b) for the cosinus covariate effect and in Figure 4.17(c) where both dynamic effects are found.



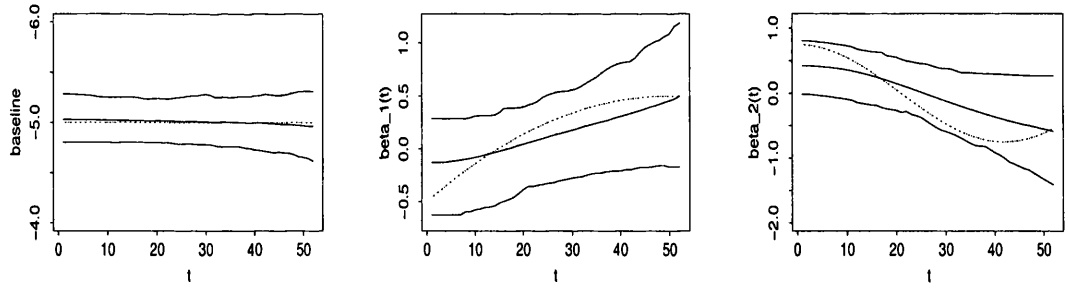
(a)

Figure 4.14: *Mean of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The drop out probability is* 3% *at each time interval* t *to* $t + 1$. *The covariate effects are* $\beta_1(t) = 1$ *and* $\beta_2(t) = -1$, *with a baseline hazard of* $\beta_0(t) = -5$. *True functions are given by the dashed lines.*



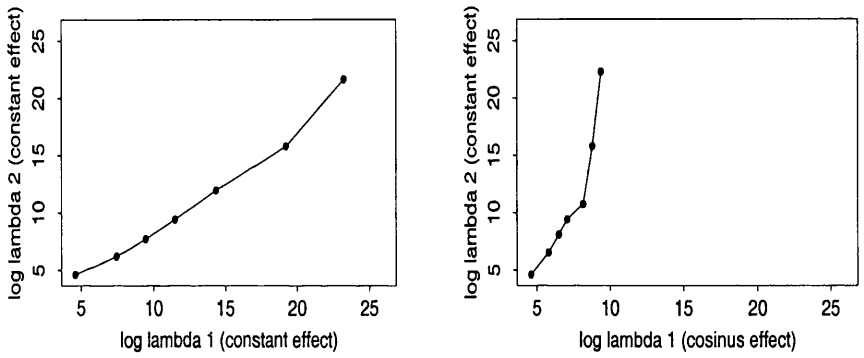
(a)

Figure 4.15: *Mean of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The drop out probability is* 3% *at each time interval* t *to* $t+1$. *The covariate effects are* $\beta_1(t)$ *as a cosinus effect and* $\beta_2(t) = 1$, *with a baseline hazard of* $\beta_0(t) = -5$. *True functions are given by the dashed lines.*

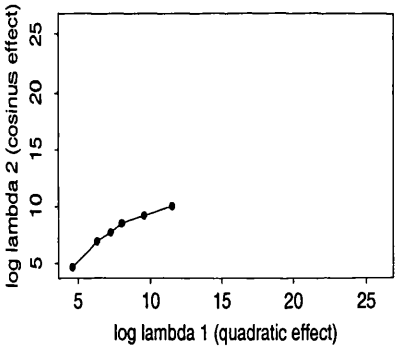


(a)

Figure 4.16: *Mean of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The drop out probability is* 3% *at each time interval* t *to* $t+1$. *The covariate effects are* $\beta_1(t)$ *as a quadratic effect and* $\beta_2(t)$ *as a cosinus effect, with a baseline hazard of* $\beta_0(t) = -5$. *True functions are given by the dashed lines.*



(a) (b)



(c)

Figure 4.17: *Steps of the hybrid smoothing parameter selection procedure starting with* $\hat{\lambda}_l^{(0)} = 100$, $l = 0, 1, 2$ *(and shown here for* $l = 1, 2$*). The filled circles indicate the steps of the smoothing parameter estimate. Plot corresponds to the smoothing parameter updates, for one simulation, (a) when both covariate effects are constant, (b) when one covariate effect is dynamic and the other is constant, (c) when both covariate effects are dynamic.*

4.7 Chapter Summary

Covariate effects in the Cox proportional hazards model are assumed to be constant over time. The use of a smooth dynamic hazard model has been demonstrated allowing estimation of smooth dynamic covariate effects in multiple covariate survival data. Working directly with the likelihood function allows a smooth estimate of the baseline hazard to be obtained. Non-proportional hazard functions were fitted in a numerically handy way using Poisson regression resulting from numerical integration of the cumulative hazard function. A hybrid smoothing parameter selection method was carried out by utilising the link between P-splines and penalised quasi-likelihood estimation in generalised linear mixed models and controlling the resulting estimates with the Akaike criterion. Finally simulations demonstrated the performance of this routine. In the next Chapter a numerically faster approach to smooth estimation of dynamic covariate effects, based on the partial likelihood function, is considered.

Chapter 5

A Partial Likelihood Approach to the Smooth Estimation of Dynamic Covariate Effects

5.1 Introduction

Smooth estimation of dynamic covariate effects based on the likelihood function was discussed in the previous Chapter. Here a smooth approach based on the partial likelihood function is considered. Smooth estimates of dynamic covariate effects are obtained using P-spline smoothing. This partial likelihood approach is numerically faster and allows for estimation of dynamic covariate effects in large survival data sets. This work can be sen as a follow on to work carried out by several authors including Gray (1994).

5.2 Smooth Partial Likelihood Estimation

Covariate effects in the Cox proportional hazards model may be estimated using the partial likelihood function. The partial likelihood function (see Section 2.4.2) can be written in its standard form as

$$L(\beta) = \prod_{i=1}^{N} \left\{ \frac{\exp(\beta^T x_i)}{\sum_{l \in R(t_i)} \exp(\beta^T x_l)} \right\}^{\delta_i},$$

where x_i is a set of covariates or risk factors and β is the vector of unknown covariate effects. In Section 2.7 the Cox proportional hazards model was extended to include dynamic effects. The dynamic Cox model is expressed as

$$h(t|\beta, x_i) = h_0(t) \exp(\beta^T(t) x_i).$$

The idea is to smoothly estimate $\beta(t)$, a vector of covariate effects varying smoothly with time using P-splines. A Poisson type approach based on the likelihood function was described in Chapter 4. This Chapter describes a numerically faster approach based on the partial likelihood function.

Let T_j denote the survival time of the jth individual and let C_j be the corresponding right censored time, $j = 1, \ldots, N$. Observe Y_j such that $Y_j = \min(T_j, C_j)$ and define the censoring indicator as $\delta_j = 1$ if $T_j < C_j$ and $\delta_j = 0$ otherwise. The observed failure point times are denoted by $t_1, \ldots, t_n$. Recall that $R(t_i)$ is the risk set at time t_i. For simplicity of notation this is written as R_i where $R_i = \{j : Y_j \geq t_i\}$. Finally define $\mathcal{D}_i$ as the index set of units failing at the time point t_i, i.e. $\mathcal{D}_i = \{j : Y_j = t_i \text{ and } \delta_j = 1\}$, and allow covariate effects β to vary with time t. The partial log likelihood for the smooth model is then defined

as

$$l(\beta(t)) = \sum_{i=1}^{n} \left[\left(\sum_{j \in \mathcal{D}_i} \beta^T(t) x_j \right) - |\mathcal{D}_i| \log \left\{ \sum_{j \in R_i} \exp(\beta^T(t) x_j) \right\} \right], \tag{5.1}$$

where $\beta(t)$ is a smooth but unknown function in time. Note that if $\beta(t)$ is constant, i.e. $\beta(t) = \beta$, then the smooth partial log likelihood in Equation (5.1) simplifies to the partial log likelihood in the case of a proportional hazards model.

Recall from Section 4.2 that when the covariate effects are constant over time, the cumulative hazard function factorises to the covariate effects multiplied by the cumulative baseline hazard and the resulting profile likelihood for the parameters is equivalent to the partial likelihood (Cox, 1972). This justification of the partial likelihood is due to Breslow (1972). If, however, covariate effects do vary with time then such factorisation of the cumulative hazard does not exist. The idea is to pretend that $\beta(t) = \beta$ and use the partial likelihood for estimation, leaving the cumulative hazard unspecified but estimating $\beta(t)$ smoothly. For kernel type smoothing this approach has been investigated theoretically in Cai and Sun (2002).

Smooth estimation of $\beta(t)$ is achieved here using penalised spline regression. Let $B(t) = \{b_1(t), \ldots, b_q(t)\}$ be a high dimensional basis developed over knots $t_1, \ldots, t_q$. It is useful to extract the intercept from the smooth function. Therefore, for the lth covariate in the model, $l = 1, \ldots, p$, this gives

$$\beta_l(t) = \beta_{0l} + \tilde{B}(t)\alpha_l,$$

where β_{0l} is the constant part, $\tilde{B}(t)$ is the $q-1$ dimensional basis matrix containing no intercept and α_l are the spline basis coefficients. As a starting value for β_0 take the estimate obtained from the Cox-PH function and set $\alpha = 0$. Using the Kro-

necker product one can write $\beta(t) = \mathbf{W}(t)\boldsymbol{\theta}$, with $\mathbf{W}(t) = I_p \otimes \{1, \tilde{B}(t)\}$ where I_p is the p dimensional identity matrix, and with parameter vector $\boldsymbol{\theta} = (\theta_1^T, \ldots, \theta_p^T)$ where $\theta_l = (\beta_{0l}, \alpha_l^T)^T$. One can use different spline bases for each of the separate components of $\beta(t)$, but for simplicity of presentation this generalisation is ignored here. The spline coefficients α_l are penalised leading to the following partial log likelihood

$$l^P(\boldsymbol{\theta}, \lambda) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) - \frac{1}{2}\sum_{l=1}^{p} \lambda_l \alpha_l^T D_l \alpha_l, \tag{5.2}$$

with

$$l_i(\boldsymbol{\theta}) = \sum_{j \in \mathcal{D}_i} x_j \mathbf{W}(t)\boldsymbol{\theta} - |\mathcal{D}_i| \log \left\{ \sum_{j \in R_i} \exp(x_j \mathbf{W}(t)\boldsymbol{\theta}) \right\}$$

as the partial likelihood contribution. Component-wise smoothing parameters λ_l, $l = 1, \ldots, p$, steer the amount of penalisation and D_l is an appropriately chosen penalty matrix (see Section 4.3.1). For notational convenience the penalty component can be rewritten as $\boldsymbol{\theta}^T(\boldsymbol{\Lambda}\mathbf{D})\boldsymbol{\theta}$ where $\mathbf{D}$ is a block diagonal matrix built from matrices diag$(0, D_l)$, with diag$(0, D_l)$ the q dimensional diagonal basis which has D_l in the bottom right hand corner and 0 elsewhere. Similarly the bandwidth matrix $\boldsymbol{\Lambda}$ is a diagonal matrix with $(\lambda_1 \otimes 1_q^T, \ldots, \lambda_p \otimes 1_q^T)$ as diagonal elements, and with 1_q as the q dimensional unit vector.

Let $s_i(\theta)$ denote the score contribution. Differentiating Equation (5.2) with respect to $\boldsymbol{\theta}$ leads to the penalised score equation

$$\frac{\partial l^P(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\mathbf{D}\theta,$$

where

$$s_i(\boldsymbol{\theta}) = \sum_{j \in \mathcal{D}_i} \mathbf{W}(t_i)^T x_j^T - |\mathcal{D}_i| \sum_{j \in R_i} \mathbf{W}(t_i)^T x_j^T \pi(j|R_i, \boldsymbol{\theta}).$$

Here $\pi(j|R_i, \theta) = \exp(x_j \mathbf{W}(t_i)\boldsymbol{\theta}) / \sum_{k \in R_i} \exp(x_k \mathbf{W}(t_i)\boldsymbol{\theta})$ are weights which sum to 1. Similarly, denote the second order derivative by $\nabla s_i(\boldsymbol{\theta})$. One can then write the penalised second order derivative with respect to $\boldsymbol{\theta}$ as

$$\frac{\partial^2 l^P(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda} \mathbf{D},$$

where

$$\begin{aligned} \nabla s_i(\boldsymbol{\theta}) \;=\; & -|\mathcal{D}_i| \left\{ \sum_{j \in R_i} \mathbf{W}(t_i)^T x_j^T x_j \mathbf{W}(t_i) \pi(j|R_i, \boldsymbol{\theta}) \right. \\ & \left. + \left(\sum_{j \in R_i} \mathbf{W}(t_i)^T x_j^T \pi(j|R_i, \boldsymbol{\theta}) \right) \left(\sum_{j \in R_i} x_j \mathbf{W}(t_i) \pi(j|R_i, \boldsymbol{\theta}) \right) \right\}. \end{aligned}$$

The variance estimates are approximated using the sandwich variance estimator

$$\text{var}(\hat{\boldsymbol{\theta}}) = - \left\{ \sum_{i=1}^{n} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda} \mathbf{D} \right\}^{-1} \left\{ \sum_{i=1}^{n} \nabla s_i(\boldsymbol{\theta}) \right\} \left\{ \sum_{i=1}^{n} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda} \mathbf{D} \right\}^{-1}.$$

For the lth fitted component, $l = 1, \ldots, p$, the variance estimator for the functional shape is obtained as $\text{Var}(\hat{\beta}_l(t)) = \text{Var}(W_l(t)\hat{\boldsymbol{\theta}}_l)$, which is equivalent to $\text{Var}(\hat{\beta}_l(t)) = W_l(t)\text{Var}(\hat{\theta}_l)W_l^T(t)$, and a pointwise 95% confidence interval for the lth component is given by

$$\text{C.I.}(\hat{\beta}_l(t)) = \hat{\beta}_l(t) \pm 1.96\sqrt{\text{Var}(\hat{\beta}_l(t))}.$$

5.3 Link to Generalised Linear Mixed Models

5.3.1 Penalised Quasi-Likelihood Estimation

The connection between P-spline smoothing and generalised linear mixed models (GLMMs) is used to choose appropriate smoothing parameters λ_l, $l = 1, \ldots, p$. Consider coefficients α_l as independent normally distributed variables with

$$\alpha_l \sim N(0, \lambda_l^{-1} D_l^-), \tag{5.3}$$

where D_l^- is the generalised inverse of D_l, $l = 1, \ldots, p$. Smoothing parameters λ_l now occur in the *a priori* variance of α_l. Interpret $\exp\left(\sum_{i=1}^n l_i(\boldsymbol{\theta})\right)$ as the partial likelihood given random coefficients α_l, where $\boldsymbol{\theta}$ is composed of β_{0l} and α_l. The partial likelihood for parameters β_{0l} and λ_l is then obtained by integrating out the random coefficients such that

$$l(\beta_{0l}, \lambda_l) = \int \exp\left(\sum_{i=1}^n l_i(\boldsymbol{\theta})\right) \prod_{l=1}^p \phi(\alpha_l, \lambda_l^{-1} D_l^-) d\alpha_l, \tag{5.4}$$

with $\phi(\cdot)$ as normal density function. Estimation of λ_l is based on the likelihood function (5.4). Approximating the integral and inserting estimates for β_{0l} leads to the Laplace approximation for the log marginal partial likelihood

$$\begin{aligned} l^{mp}(\lambda) \approx & \sum_{i=1}^n l_i(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \sum_{l=1}^p (\lambda_l \hat{\alpha}_l^T D_l \hat{\alpha}_l + \log|\lambda_l D_l|) \\ & - \frac{1}{2} \log \left| \sum_{i=1}^n \nabla s_i(\hat{\boldsymbol{\theta}}) + \operatorname{diag}(\lambda_l D_l) \right|, \end{aligned} \tag{5.5}$$

with $\lambda = (\lambda_1, \ldots, \lambda_p)$ and with $\hat{\boldsymbol{\theta}}$ maximising the right hand side of Equation (5.5). Ignoring the last component in Equation (5.5), by assuming that it depends only weakly on $\hat{\boldsymbol{\theta}}$ (for the same argument see Breslow and Clayton, 1993), $\hat{\boldsymbol{\theta}}$ is the penalised estimate obtained by maximising Equation (5.2). Differentiating Equation (5.5) with respect to λ_l leads to

$$0 = \hat{\alpha}_l^T D_l \hat{\alpha}_l + \frac{q}{\lambda_l} - tr\left(\left(\sum_{i=1}^{n} \nabla s_i(\hat{\boldsymbol{\theta}}) + \text{diag}(\lambda_l D_l)\right)^{-l} D_l\right), \tag{5.6}$$

where $()^{-l}$ refers to the lth diagonal block component of the inverse penalised second derivative. For numerical convenience, let $\left(\sum_{i=1}^{n} \nabla s_i(\hat{\boldsymbol{\theta}}) + \text{diag}(\lambda_l D_l)\right)^{-l}$ be approximated by $\left\{\left(\sum_{i=1}^{n} \nabla s_i(\hat{\boldsymbol{\theta}})\right)^{l} + \lambda_l D_l\right\}^{-1}$. This allows Equation (5.6) to be simplified leading to the penalised quasi-likelihood estimate

$$\hat{\lambda}_l = \frac{\text{df}_l}{\hat{\alpha}_l^T D_l \hat{\alpha}_l} \tag{5.7}$$

with df_l as the approximate degrees of freedom for the lth smooth component as defined by

$$\text{df}_l = \text{tr}\left\{\left(\sum_{i=1}^{n} \nabla s_i(\hat{\boldsymbol{\theta}}) - \boldsymbol{\Lambda}\mathbf{D}\right)^{-l}\left(\sum_{i=1}^{n} \nabla s_i(\hat{\boldsymbol{\theta}})\right)^{l}\right\}.$$

It is worth pointing out that Equation (5.7) is not an explicit solution since estimate $\hat{\lambda}_l$ depends on estimates $\hat{\boldsymbol{\theta}}$ and vice versa. One can however update λ by cycling between estimation of $\boldsymbol{\theta}$ for given λ and estimation of λ for given $\boldsymbol{\theta}$.

5.3.2 Hybrid Smoothing Parameter Selection

The link to linear mixed models relates the smoothing parameters λ_l, $l = 1, \ldots, p$, to variance components in a mixed model. The smoothing parameters can then

be estimated by an approximate likelihood, leading to (5.7). For normal response models, the connection between the different roles of λ_l in the smooth and mixed model, respectively, is investigated in Kauermann (2004). As shown there, even though the connection and the simple form (5.7) is elegant, the estimation of a smoothing parameter λ via (5.7) uncovers problems in that the resulting smooth estimated fit of $\hat{\beta}_l(t)$ is not optimal in a Mean Squared Error sense. For this reason, the intention is to find a smoothing parameter λ which provides good properties with respect to its bias-variance trade off. As a suggestion therefore λ is estimated in a hybrid way by taking advantage of the numerical simplicity of the PQL estimate given in (5.7) but controlling smoothing parameters using the Akaike information criterion. Therefore at the jth cycle of the PQL estimation the Akaike criterion $\mathrm{AIC}(\hat{\lambda}^{(j)})$ is calculated with

$$\mathrm{AIC}(\lambda) = -2\sum_{i=1}^{n} l_i(\hat{\boldsymbol{\theta}}) + 2\mathrm{df}(\lambda),$$

where $\mathrm{df} = \sum_{l=1}^{p} \mathrm{df}_l$ is the degrees of freedom of the model. This allows the value of the Akaike information criterion to be checked at each iteration and iterations are stopped if $\mathrm{AIC}(\hat{\lambda}^{(j+1)}) > \mathrm{AIC}(\hat{\lambda}^{(j)})$. In this way, an estimated smoothing parameter $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ is obtained which provides a small (though not minimal) value of the Akaike information criterion. At the final estimate $\hat{\lambda}^{(j)}$ a local grid search could now be run to find the minimum of $\mathrm{AIC}(\lambda)$. For ease of numerical effort, however, this step is omitted.

5.4 Simulations

5.4.1 Single Covariate

This Chapter describes the use of P-spline smoothing for estimating dynamic covariate effects in hazard models, within a partial likelihood setting. Simulations will now be used to demonstrate the performance of this approach. First consider simulating survival data with a single covariate. Survival data are simulated on a discrete time grid, $t = 1, 2, \ldots, 60$. For each simulation, a binary covariate is randomly chosen such that $P(x_1 = 1) = 0.5$. The baseline hazard function is constant, $\beta_0(t) = -5$. Drop out probabilities are either 3% or 0.1% at each time interval t to $t + 1$, with the sample size being either $N = 400$ or $N = 4000$. In the first simulation setting the covariate effect is kept constant while in following settings the covariate effect is one of the dynamic effects $\beta(t)$ described in Table 4.1 and shown in Figure 4.2. For a constant covariate effect, $\beta(t) = 1$, Figure 5.1 shows the estimated survivor function (left plot) and separate Kaplan-Meier estimates of the survivor function for individuals within each group (right plot) where group 0 corresponds to the baseline group. Plots correspond to a sample size of $N = 4000$. Similar plots for $N = 400$ were shown in Figure 4.3 in the previous Chapter. Estimates are based on a fitted Cox proportional hazards model.

Linear truncated splines are used with the number of knots q chosen in line with the guidelines given in Ruppert et al. (2003). Smoothing parameter selection is started with $\hat{\lambda}_l^{(0)} = 100$, for $l = 1$, and is updated with $\lambda_l^{(t)}$ as long as the Akaike criterion decreases. Shown in Figure 5.2 are the mean estimates from 150

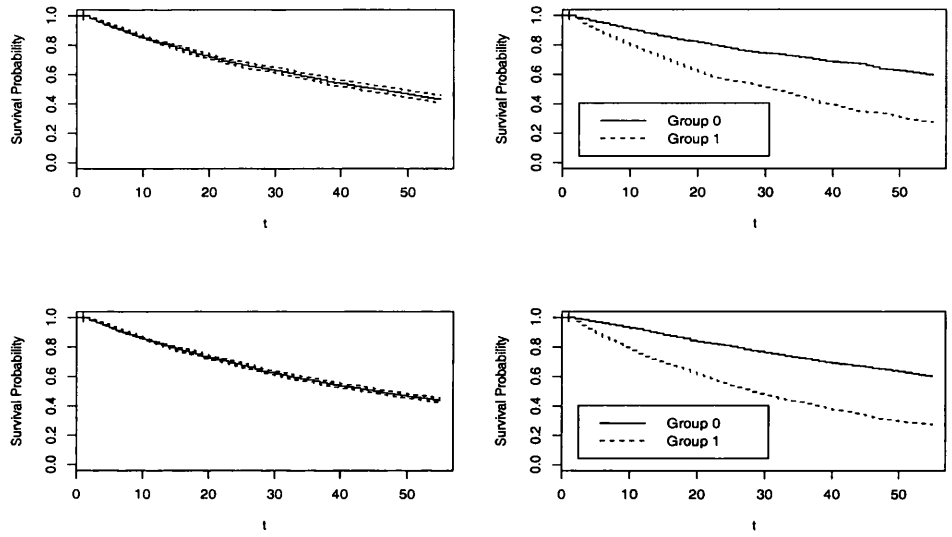


Figure 5.1: *Kaplan-Meier estimates (left plots) and Kaplan-Meier estimates by status (right plots) corresponding to a sample size of* $N = 4000$. *In the top row, the drop out probability is* 3% *at each time interval* t *to* $t + 1$, *in the bottom row, the drop out probability is* 0.1% *at each time interval* t *to* $t + 1$.

simulations together with corresponding pointwise 90% confidence intervals for a time constant effect. The top row shows those simulations corresponding to a sample size of $N = 400$, the bottom row corresponding to $N = 4000$. The left hand plots shows simulations corresponding to a drop out probability of 3% at each time interval t to $t + 1$ and the right hand plots to a drop out probability of 0.1% at each time interval t to $t + 1$.

Simulated coverage probabilities are shown in Figures 5.7 to 5.11. One can see from Figure 5.7 that simulated coverage probabilities for the constant effect are high regardless of the sample size and drop out probability. Hence a time constant fit is well captured. For the dynamic effects, shown in Figures 5.8 to 5.11, simulated coverage probabilities vary. Generally they are best when $N = 4000$ and when the drop out probability is low.

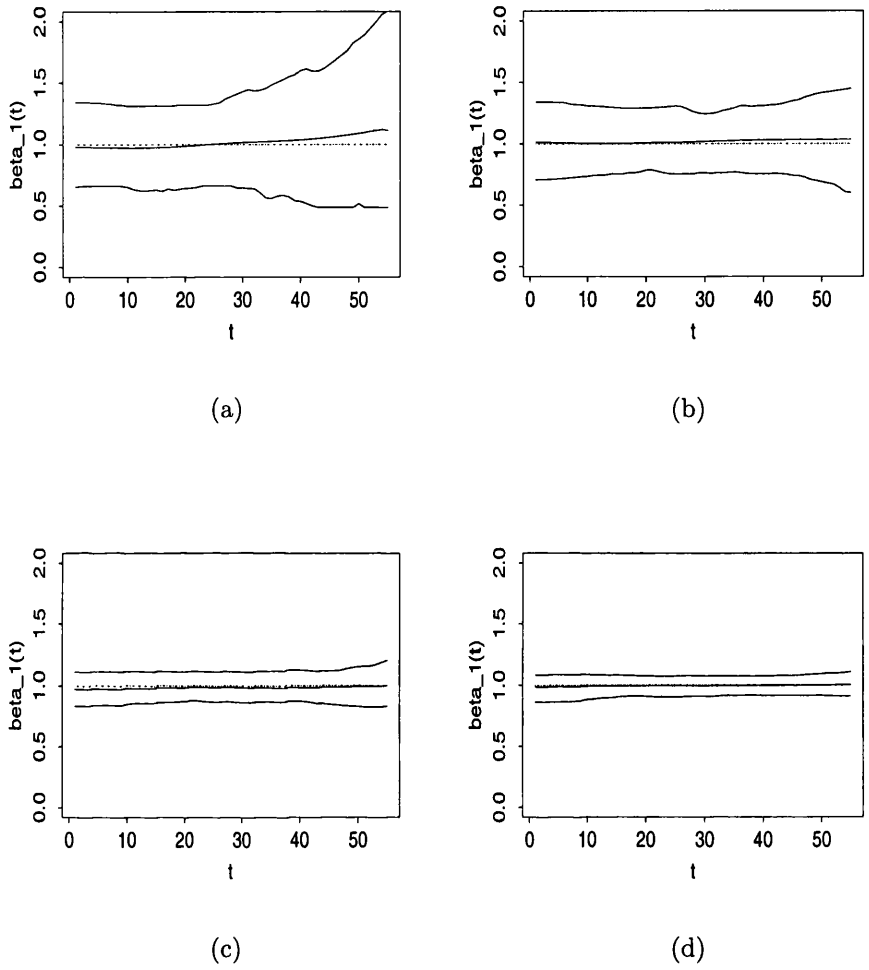


Figure 5.2: *P-spline partial likelihood estimates of a constant covariate effect ($\beta(t) = 1$). Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row corresponds to* $N = 400$ *with* (a) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (b) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The bottom row corresponds to* $N = 4000$ *with* (c) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (d) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The true function is given by the dashed line.*

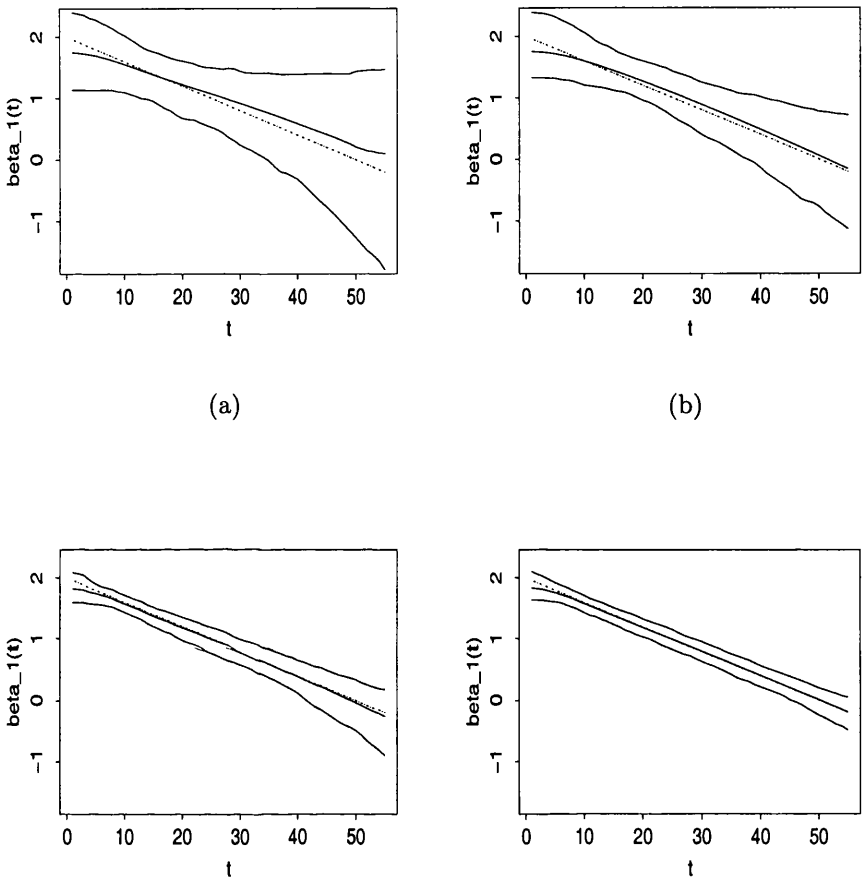


Figure 5.3: *P-spline partial likelihood estimates of a linear (steep) covariate effect. Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row corresponds to* $N = 400$ *with* (a) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (b) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The bottom row corresponds to* $N = 4000$ *with* (c) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (d) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The true function is given by the dashed line.*

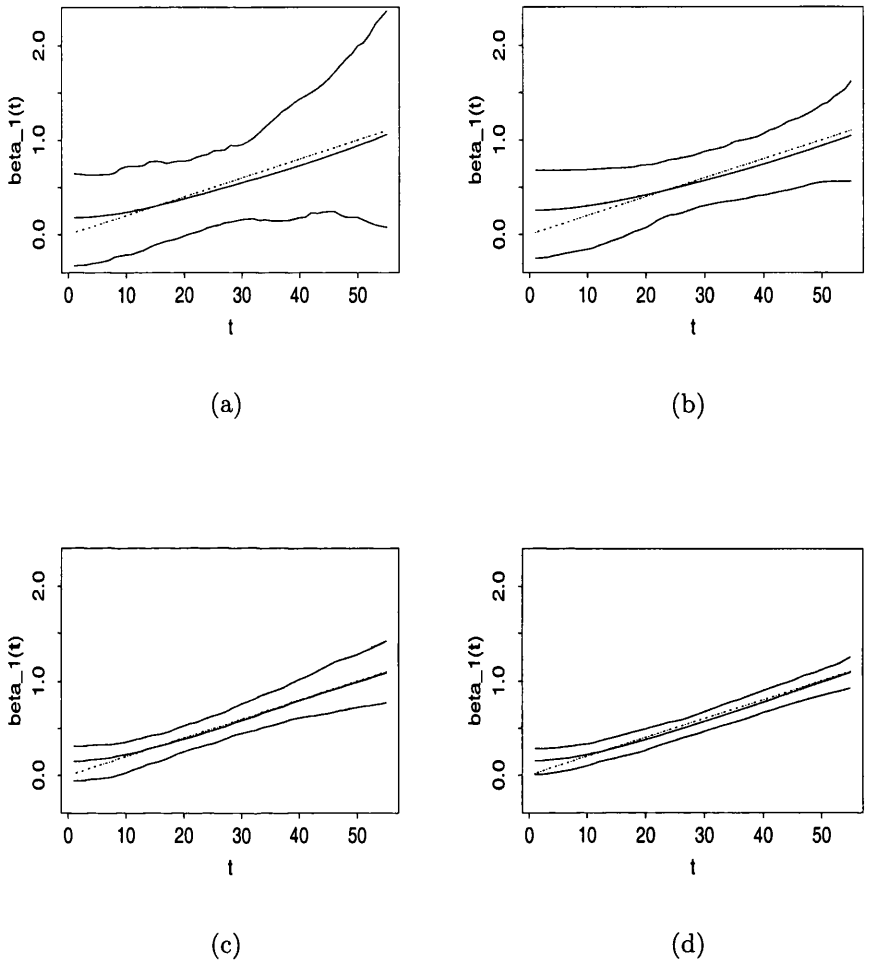


Figure 5.4: *P-spline partial likelihood estimates of a linear covariate effect. Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row corresponds to* $N = 400$ *with* (a) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (b) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The bottom row corresponds to* $N = 4000$ *with* (c) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (d) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The true function is given by the dashed line.*

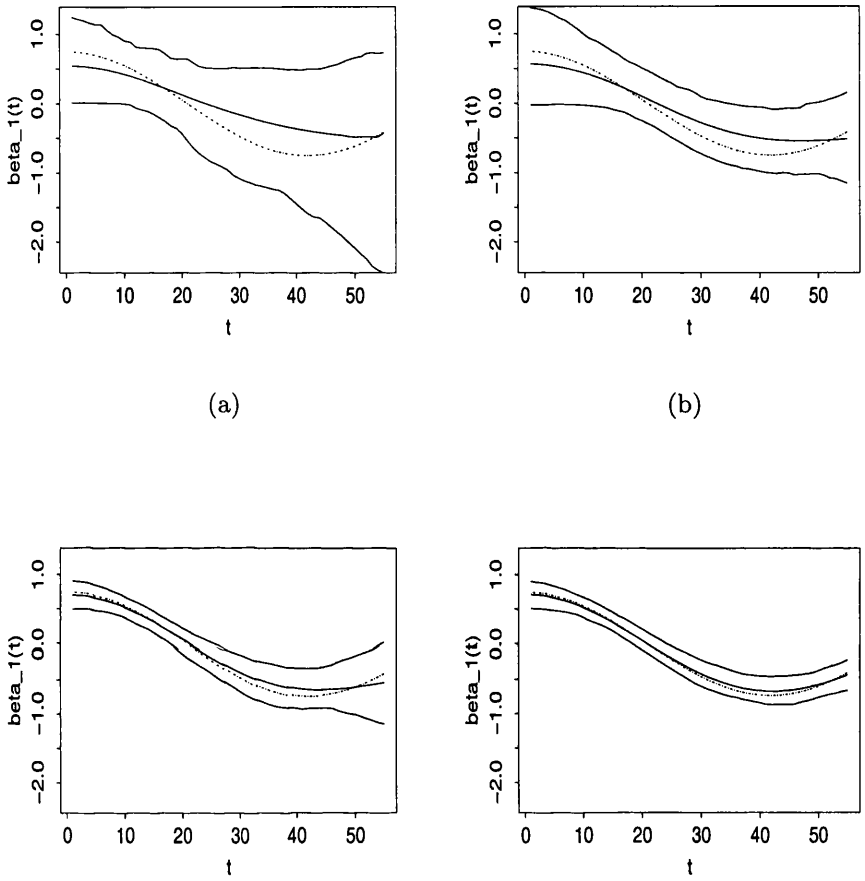


Figure 5.5: *P-spline partial likelihood estimates of a cosinus covariate effect. Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row corresponds to* $N = 400$ *with* (a) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (b) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The bottom row corresponds to* $N = 4000$ *with* (c) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (d) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The true function is given by the dashed line.*

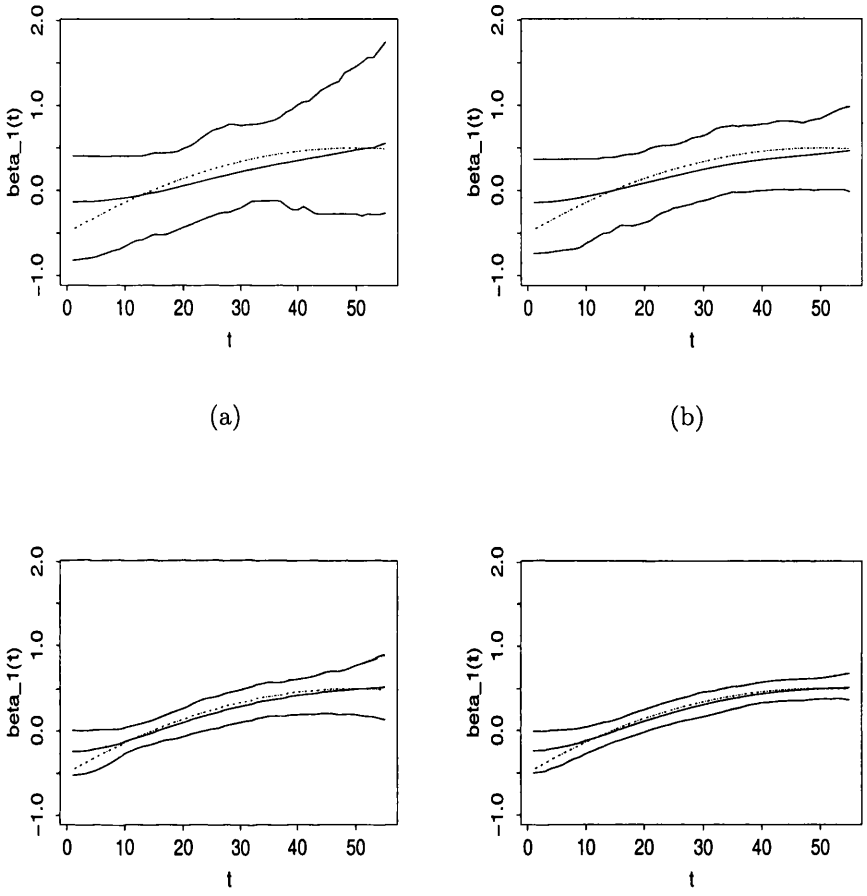


Figure 5.6: *P-spline partial likelihood estimates of a quadratic covariate effect. Shown are the means of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The top row corresponds to* $N = 400$ *with* (a) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (b) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The bottom row corresponds to* $N = 4000$ *with* (c) *a drop out probability of* 3% *at each time interval* t *to* $t+1$ *and* (d) *a drop out probability of* 0.1% *at each time interval* t *to* $t+1$. *The true function is given by the dashed line.*

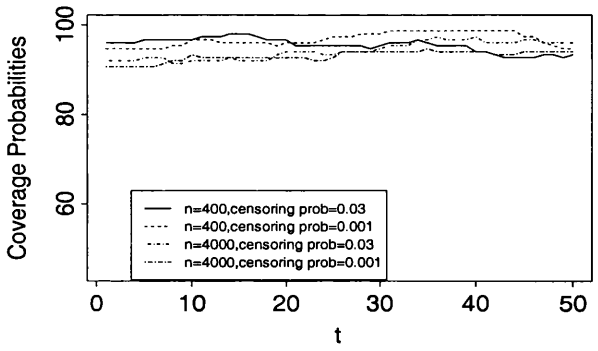


Figure 5.7: *Plot of coverage probabilities when the covariate effect is constant. Nominal value is* 95%.

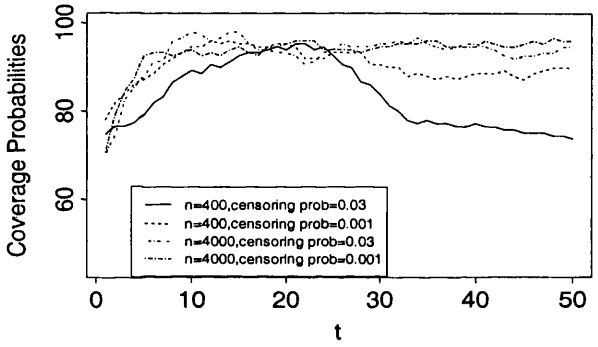


Figure 5.8: *Plot of coverage probabilities when the covariate effect is linear (steep). Nominal value is* 95%.

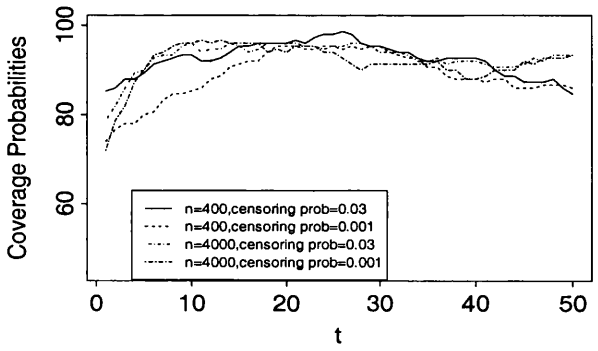


Figure 5.9: *Plot of coverage probabilities when the covariate effect is linear. Nominal value is* 95%.

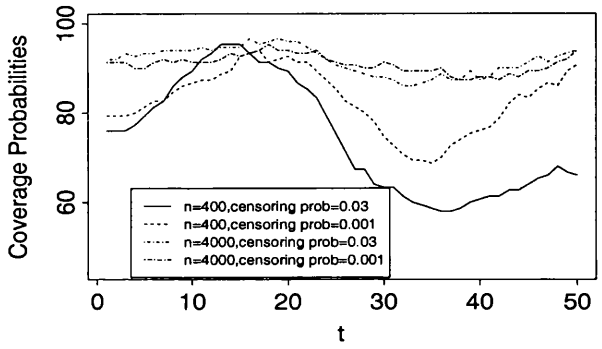


Figure 5.10: *Plot of coverage probabilities when the covariate effect is cosinus. Nominal value is* 95%.

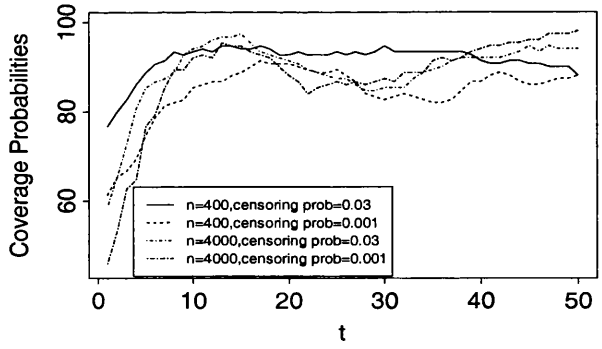


Figure 5.11: *Plot of coverage probabilities when the covariate effect is quadratic. Nominal value is* 95%.

5.4.2 Multiple Covariates

Now consider survival data simulated with two covariates on a discrete time grid, $t = 1, 2, \ldots, 60$. Sample sizes are either $N = 400$ or $N = 4000$ with a drop out probability of 3% at each time interval t to $t + 1$. The two covariates x_1 and x_2 are randomly chosen with $P(x_1 = 1) = 0.5$ and $P(x_2 = 1) = 0.3$. The baseline hazard, $\beta_0(t) = -5$, is kept constant. Smoothing parameter selection is started with $\hat{\lambda}_l^{(0)} = 100$, for $l = 1, 2$, and is updated with $\lambda_l^{(t)}$ as long as the Akaike criterion decreases. Table 4.2 shows the two covariate effects to be simulated, in each of three settings.

The means of 150 simulations with corresponding pointwise 90% confidence intervals for survival data simulated with two constant covariate effects are shown in Figure 5.12(a) for $N = 400$ and in Figure 5.12(b) for $N = 4000$. Again one can see that for larger sample sizes the mean estimates are closer to the true covariate effects and pointwise 90% confidence intervals are narrower. Figures 5.13(a) and 5.13(b) show the means and pointwise intervals for survival data simulated with one dynamic cosinus effect and one constant effect, for $N = 400$ and $N = 4000$ respectively. For survival data simulated with two dynamic effects, one can see the means and pointwise intervals from 150 simulations for $N = 400$ and $N = 4000$ in Figures 5.14(a) and 5.14(b).

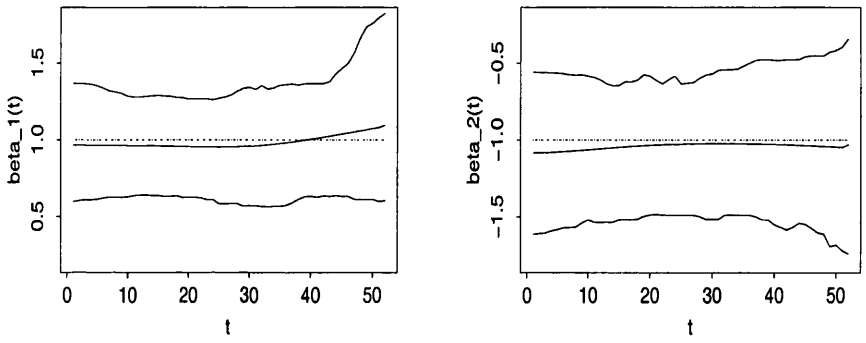
The plots in Figure 5.15 show the steps of the smoothing parameter estimate $\hat{\lambda}_l^{(j)}$, $l = 1, 2$, for one simulation when data are simulated with 2 constant effects, with 1 dynamic and 1 constant effect and with 2 dynamic effects. The smoothing parameter selection procedure is started each time with $\hat{\lambda}_l^{(0)} = 100$. Final estimates are based on the Akaike criterion and steps are terminated when

$AIC(\lambda^{(j+1)}) > AIC(\lambda^{(j)})$. Generally around five or six steps are needed to achieve the data driven smoothing parameter estimate. From these plots, one can see that the smoothing parameter update is very large with $\hat{\lambda}_l \to \infty$, $l = 1, 2$, when both covariate effects are constant meaning that proportional hazards fits are obtained. When a dynamic effect is present, the smoothing parameter estimate does not go to infinity. In Figure 5.15(b) one dynamic fit is captured while in Figure 5.15(c) both dynamic effects are captured.

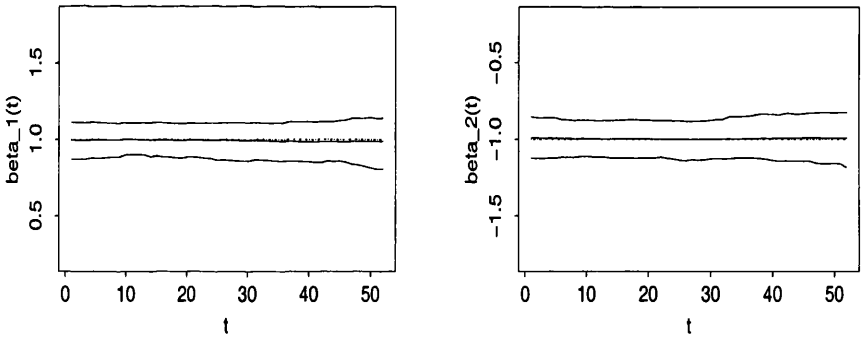
5.5 Computational Issues

As far as possible, parameters for the Poisson type approach and the partial likelihood approach are kept equal. For both approaches, survival data was simulated for $N = 400$ observations with corresponding mean estimates and pointwise 90% confidence intervals shown for 150 simulations (survival data is also simulated for $N = 4000$ in the partial likelihood case). A suggested minimum sample size, using both approaches, is $N = 100$. Satisfactory results were obtained using this as guideline, however smaller sample sizes are problematic in terms of dynamic effects not being picked up. In both cases linear truncated splines were used with q, the dimension of the basis, chosen in line with the guidelines given in Ruppert et al. (2003) (for the Poisson type approach, consideration was also given to the number of grid points K, chosen to be equal to q).

Both approaches captured a constant covariate effect well, with high coverage probabilities. There was evidence with both approaches that some of the dynamic covariate effects were over smoothed. This is likely to be due to the small sample

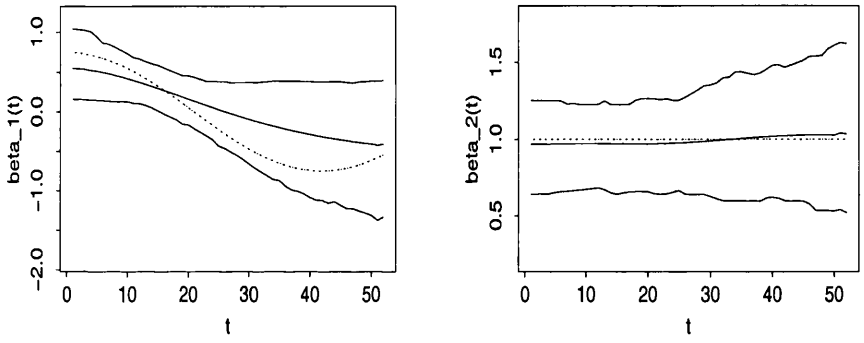


(a)

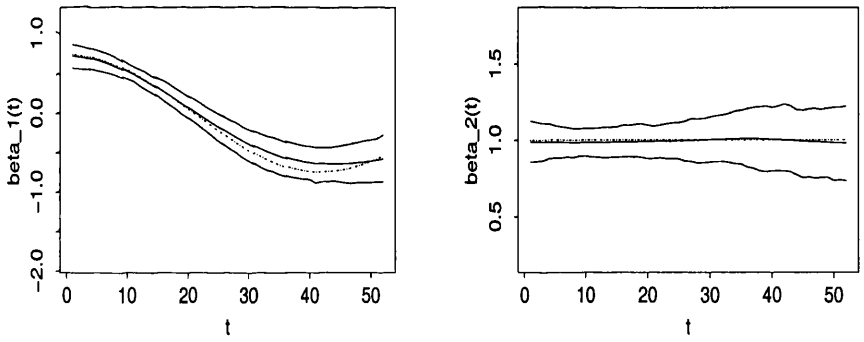


(b)

Figure 5.12: *Mean of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The drop out probability is* 3% *at each time interval* t *to* $t+1$. *The covariate effects are* $\beta_1(t) = 1$ *and* $\beta_2(t) = -1$. *Sample sizes are* $N = 400$ *in* (a) *and* $N = 4000$ *in* (b). *True functions are given by the dashed lines.*

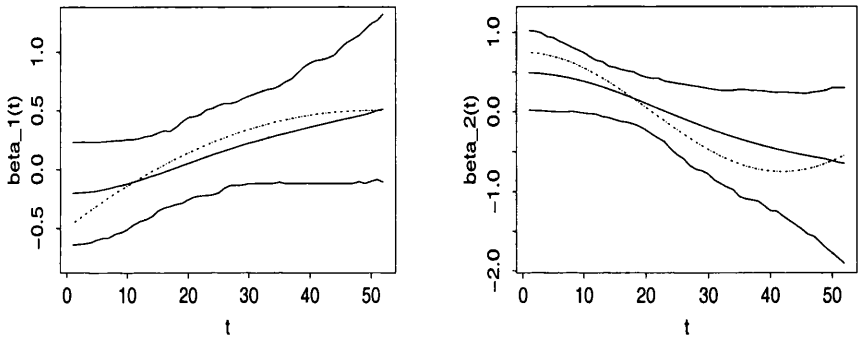


(a)

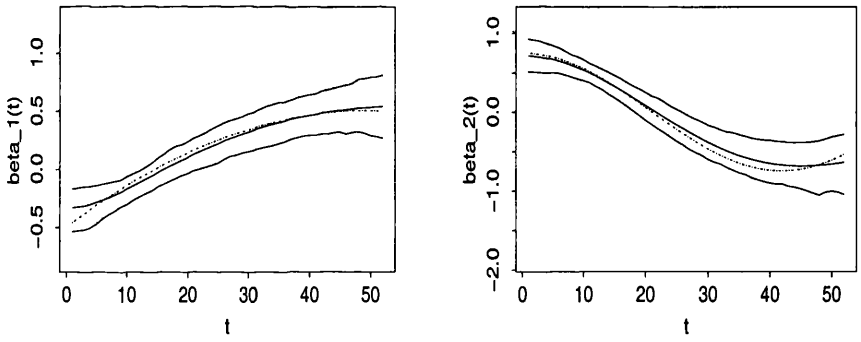


(b)

Figure 5.13: *Mean of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The drop out probability is* 3% *at each time interval* t *to* $t+1$. *The covariate effects are* $\beta_1(t)$ *as a cosinus effect and* $\beta_2(t) = 1$. *Sample sizes are* $N = 400$ *in* (a) *and* $N = 4000$ *in* (b). *True functions are given by the dashed lines.*

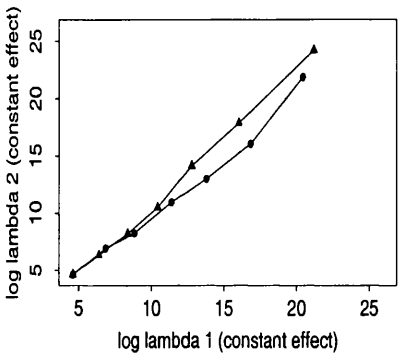


(a)

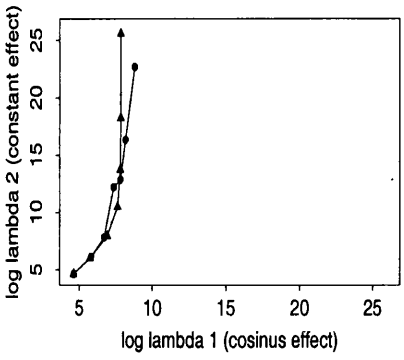


(b)

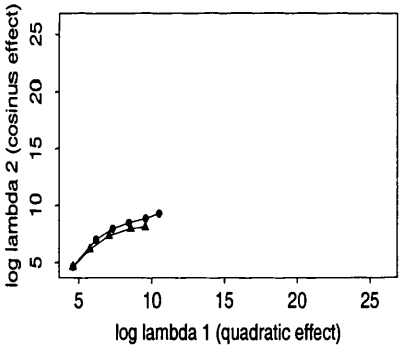
Figure 5.14: *Mean of* 150 *simulations with corresponding pointwise* 90% *confidence intervals. The drop out probability is* 3% *at each time interval* t *to* $t+1$. *The covariate effects are* $\beta_1(t)$ *as a quadratic effect and* $\beta_2(t)$ *as a cosinus effect. Sample sizes are* $N = 400$ *in* (a) *and* $N = 4000$ *in* (b). *True functions are given by the dashed lines.*



(a)



(b)



(c)

Figure 5.15: *Steps of the hybrid smoothing parameter selection procedure starting with* $\hat{\lambda}_l^{(0)} = 100$, $l = 1, 2$. *The filled circles indicate the steps of the smoothing parameter estimate when* $N = 400$, *the filled triangles indicate the steps when* $N = 4000$. *Plot corresponds to the smoothing parameter updates, for one simulation, (a) when both covariate effects are constant, (b) when one covariate effect is dynamic and the other is constant, (c) when both covariate effects are dynamic.*

size. When the sample size was increased to $N = 4000$ and analysed using the partial likelihood approach, the resulting mean estimates were much closer to the true shape of the underlying dynamic effect and coverage was higher. Despite being improved, there was still problems with coverage for some of the dynamic effects.

Confidence intervals for the fits, $\hat{\beta}_l(t)$, are constructed using the sandwich variance estimator and are based on the smoothing parameter estimates, however, the uncertainty in estimated smoothing parameters has not been accounted for. This uncertainty is a possible cause of the poor coverage probabilities observed. For a general discussion of the sandwich covariance estimator see, for example, Kauermann and Carroll (2001), or more recently Wood (2004) who suggests a Bayesian approach to the construction of confidence intervals when uncertainty in the smoothing parameter estimates are unaccounted for.

Smoothing parameter selection procedure was started each time with $\hat{\lambda}_l^{(0)} = 100$, ($l = 0, 1, 2$ for the Poisson type approach and $l = 1, 2$ for the partial likelihood approach) with steps terminated when $\text{AIC}(\hat{\lambda}^{(j+1)}) > \text{AIC}(\hat{\lambda}^{(j)})$. It can be seen from Figure 4.17 and Figure 5.15 that, for both approaches, the data driven estimate is usually obtained in around five to six steps and that steps for both approaches follow a fairly similar path.

5.6 Chapter Summary

In this Chapter P-spline smoothing was used to enable smooth estimation of dynamic covariate effects while working in the partial likelihood setting. Although

no explicit baseline estimate is obtained, moving away from the Poisson type approach allows for estimation of dynamic covariate effects in larger data sets. Simulations demonstrated the performance of this routine. In the next Chapter this approach will be applied to data from the German Socio-Economic Panel and to data from the West of Scotland Coronary Prevention Study.

Chapter 6

Illustrative Data Sets

6.1 Introduction

There are two illustrative data sets used in this thesis. The first is unemployment data from the German Socio-Economic Panel (SOEP). In this data set smooth dynamic covariate effects will be investigated using both the Poisson type approach from Chapter 4 and the partial likelihood approach from Chapter 5. The second is data from the West of Scotland Coronary Prevention Study (WOSCOPS). This is a large data set containing observations from over $6,500$ participants. Smooth dynamic covariate effects in the WOSCOPS data will be investigated using the partial likelihood approach from Chapter 5.

6.2 Unemployment Data

The German Socio-Economic Panel (SOEP) is a representative longitudinal study of private households in Germany. Data is collected annually and questionnaires cover a wide range of topics, included household composition, earnings and health. (The data set is available for scientific users from the German Institute for Economic Research, see www.diw.de). Considered here is unemployment data collected between the years 1990 to 2000. Only those individuals who are in their first spell of unemployment (N = 537) during the 10 year observation period are considered for this analysis. Survival time is recorded in months. Status is recorded as 1 if the individual enters full time employment, 2 if part time employment is entered and 0 if the individual is censored (maternity leave, school, military service, retirement, etc.). There are 34 missing observations due to non-response. Individuals with missing observations were excluded from analysis. Further information collected (with breakdown %'s shown in parenthesis) is shown in Table 6.1. Also recorded is the age (in years) of the unemployed individual. In terms

Table 6.1: *Variables of the unemployment data*

Variables	Description (%)	
Sex	1: Male (43.9)	0: Female (56.1)
Nationality	1: German (84)	0: Foreigner (16)
West-East	1: West German (23.9)	0: East German (76.1)

of age, it would be of interest to compare the hazard rates for three age groups. Therefore classify as *young* if age is less than 35 (30.6%), as *middle-aged* if age is between 35–55 (34%) and as *old* if age is greater than 55 (35.4%) years old. Two dummy variables are created: age1 coded as 1 if age is between 35–55, 0

otherwise and age2 coded as 1 if age is greater than 55, 0 otherwise. In this case both age groups are compared to those younger then 35. An event occurs when an individual gains *full time or part time* employment (total number of events = 126). Table 6.2 shows the output from a fitted Cox proportional hazards model with full time or part time employment as outcome.

Table 6.2: *A Cox PH model fitted to the unemployment data with full time or part time employment as outcome*

Variables	β	p Value	Risk Ratio
Sex	0.56	0.0022	1.75[1.22, 2.50]
Nationality	0.88	0.0035	2.42[1.34, 4.38]
West-East	1.13	< 0.0001	3.09[1.99, 4.79]
Age1	-0.40	0.084	0.67[0.43, 1.06]
Age2	-0.58	0.018	0.56[0.35, 0.91]

The estimates of covariate effects β, under a proportional hazards model, are given in column 1 with p Values in column 2. Risk ratios with 95% confidence intervals are given in column 3. From Table 6.2 one can see that for sex, for example, the risk ratio is 1.75 which has a p Value of 0.0022. Hence, under the assumption of proportional hazards, males have a significantly better chance of gaining employment than females. The variables nationality, west-east and age2 are also significant meaning that native Germans are more likely to gain employment than foreigners, that those in West Germany are more likely to gain employment than those in East Germany and that younger people are more likely to gain employment than those over 55 years old. There appears to be no significant difference between middle-aged people and young people in terms of chances of gaining employment. Standard tests of significance assessing non-

proportionality show that the variable west-east has significant departures from proportionality ($p = 0.0246$).

Figure 6.1 shows the smooth penalised estimates obtained using the Poisson type approach as described in Chapter 4. One can see from this plot that the chances of gaining employment are greater for men than for women, that native Germans have more chance of gaining employment than foreigners and that young people are more likely to gain employment than those over 55 years old. Young people, however, are no more likely to gain employment than middle-aged people. All these effects remain constant over time, hence for these variables a proportional hazards fit seems reasonable. In terms of those from West and East Germany, individuals from West Germany are more likely to gain employment, however, this effect is only significant for a period of 20 months after which the effect becomes non-significant. In this case, a dynamic effect has been captured. These findings agree with standard tests of significance assessing non-proportionality. The baseline hazard increases with time meaning that as individuals remain in the job market, their chances of gaining employment improves. The penalised estimates obtained from the partial likelihood approach, as described in Chapter 5, are given in Figure 6.2. It can be seen that results are very similar to those obtained using the Poisson type approach. Possibly the only difference it that the dynamic effect for the variable west-east is more smoothed when using the Poisson type approach.

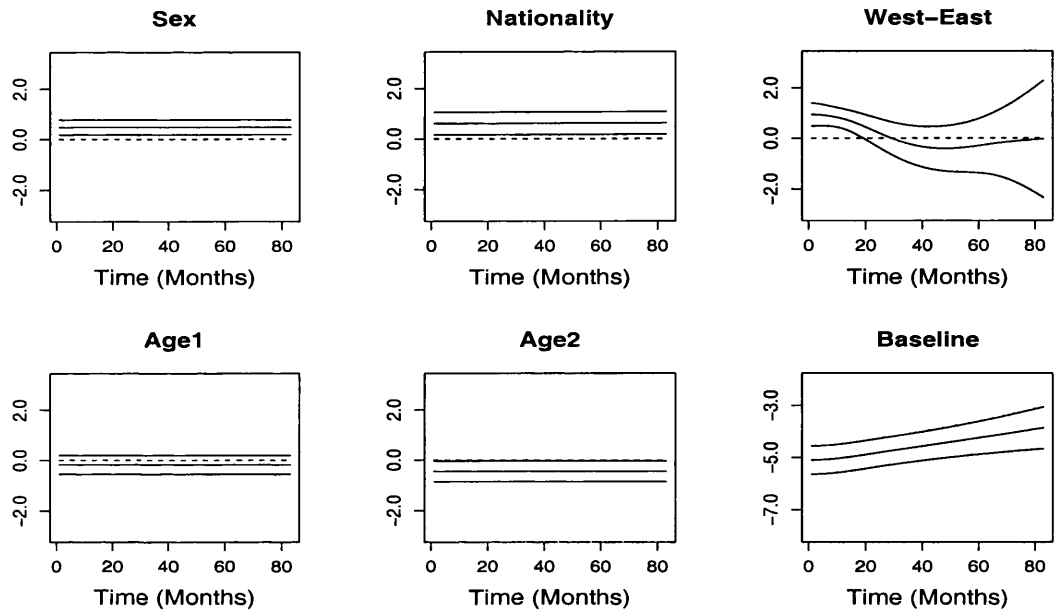


Figure 6.1: *Smooth dynamic covariate effects of the unemployment data based on the Poisson type approach. Outcome is full or part time employment. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

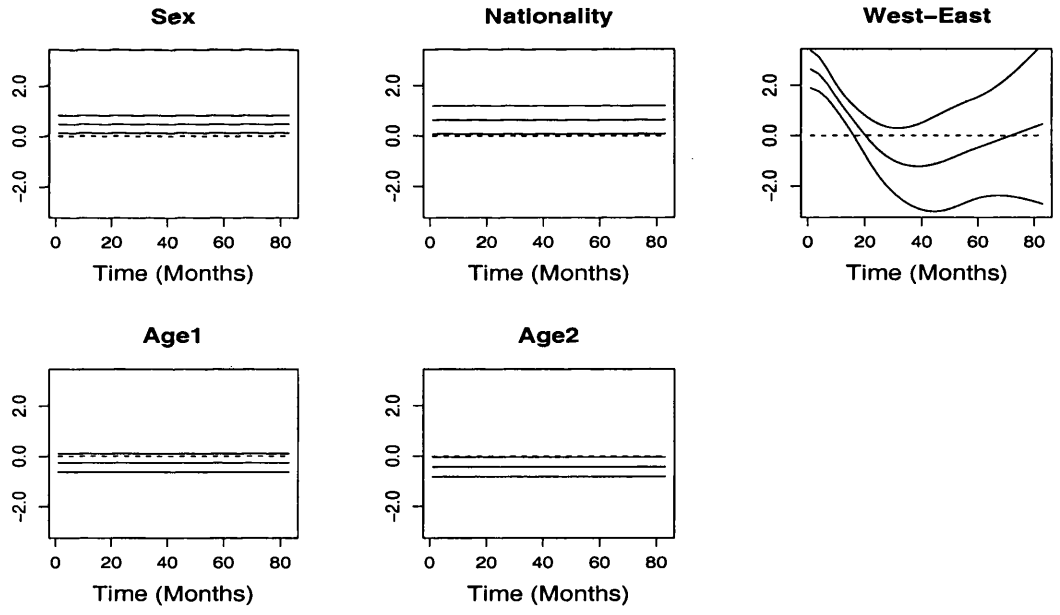


Figure 6.2: *Smooth dynamic covariate effects of the unemployment data based on the partial likelihood approach. Outcome is full or part time employment. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

6.2.1 West and East Germany

Covariate effects for West Germany and East Germany are now examined separately. In West Germany (N = 120) there were a total of 32 events while in East Germany (N = 383) a total of 94 events. Table 6.3 gives the output from a fitted Cox proportional hazards model to the West German unemployment data. The estimates of covariate effects β, under the proportional hazards assumption, are given in column 1 with p Values in column 2 and risk ratios with confidence intervals in column 3. From Table 6.3 one can see that the only significant variable is nationality. Native Germans in West Germany are more likely to gain employment than foreigners. The variables sex and age are non-significant.

Table 6.3: *A Cox PH model fitted to the West German unemployment data with full time or part time employment as outcome*

Variables	β	p Value	Risk Ratio
Sex	0.05	0.910	1.05[0.49, 2.25]
Nationality	1.13	0.043	3.08[1.04, 9.17]
Age1	−0.62	0.230	0.54[0.19, 1.48]
Age2	−0.01	0.980	0.99[0.40, 2.42]

Figures 6.3 and 6.4 show smooth penalised estimates obtained from the Poisson type approach and the partial likelihood approach respectively. These outputs are very similar and show that the only significant effect is nationality, with native Germans more likely to find employment than foreigners. All other covariate effects are non-significant. No effects were found to be dynamic with time, findings which corresponds to standard tests of significance assessing non-proportionality. The baseline hazard remains constant over time.

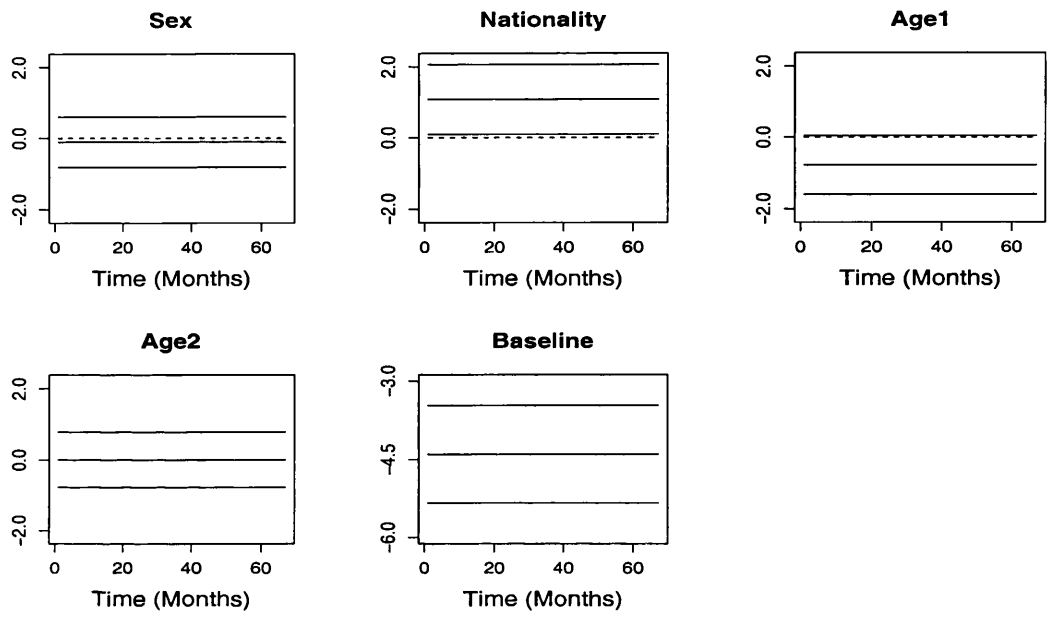


Figure 6.3: *Smooth dynamic covariate effects of the West German unemployment data based on the Poisson type approach. Outcome is full or part time employment. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

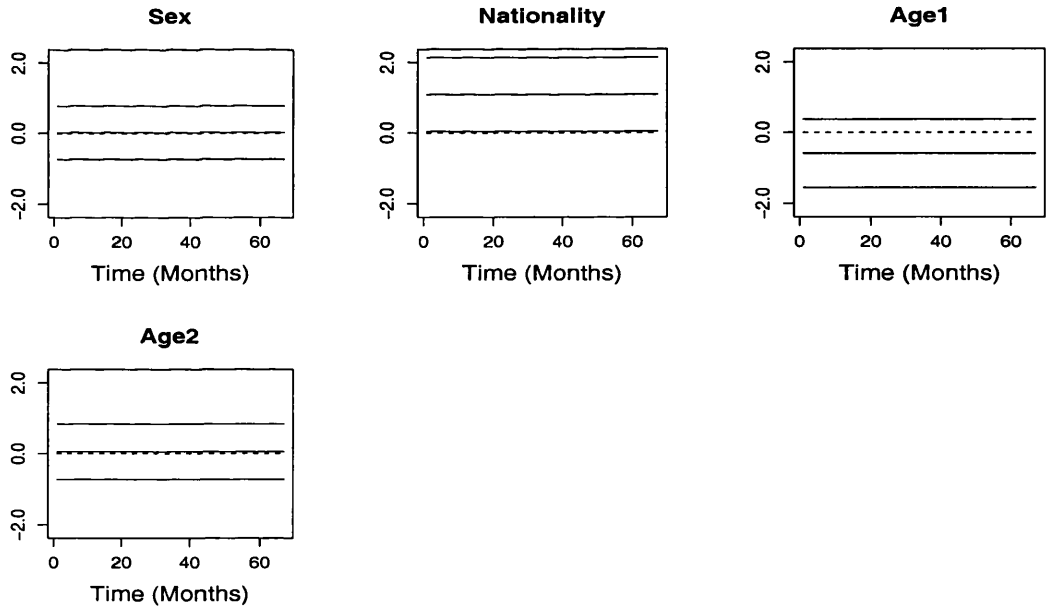


Figure 6.4: *Smooth dynamic covariate effects of the West German unemployment data based on the partial likelihood approach. Outcome is full or part time employment. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

Now consider those individuals from East Germany. The output from a fitted Cox proportional hazards output is given in Table 6.4. From this output one can see that the variables sex and age2 are significant. Under the assumption of proportional hazards nationality and age1 are non-significant.

Table 6.4: *A Cox PH model fitted to the East German unemployment data with full time or part time employment as outcome*

Variables	β	p Value	Risk Ratio
Sex	0.75	0.0005	2.12[1.39, 3.24]
Nationality	0.53	0.150	1.70[0.83, 3.48]
Age1	−0.29	0.290	0.75[0.43, 1.28]
Age2	−0.69	0.020	0.50[0.28, 0.90]

Figures 6.5 and 6.6 show the smooth penalised estimates obtained from the Poisson type approach and the partial likelihood approach respectively. Again, these outputs are very similar. They show that both sex and age2 are significant, meaning that males are more likely to find employment than females and that young people are more likely to find employment than those over 55 years of age. Both nationality and age2 are non-significant. No covariate effects were found to be dynamic with time which corresponded to findings from standard tests of significance assessing non-proportionality. The baseline hazard increases slightly over time.

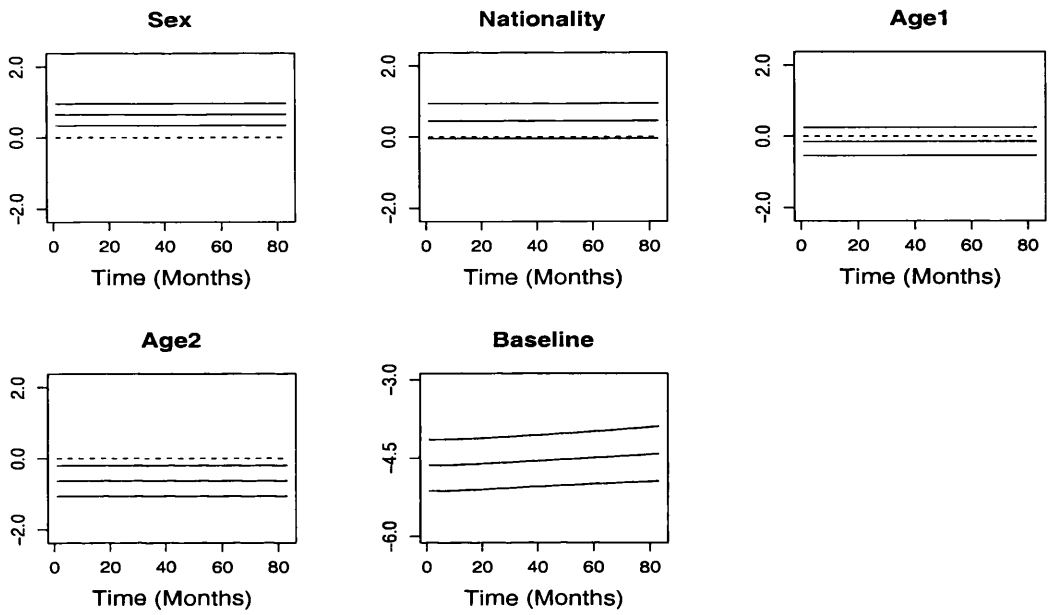


Figure 6.5: *Smooth dynamic covariate effects of the East German unemployment data based on the Poisson type approach. Outcome is full or part time employment. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

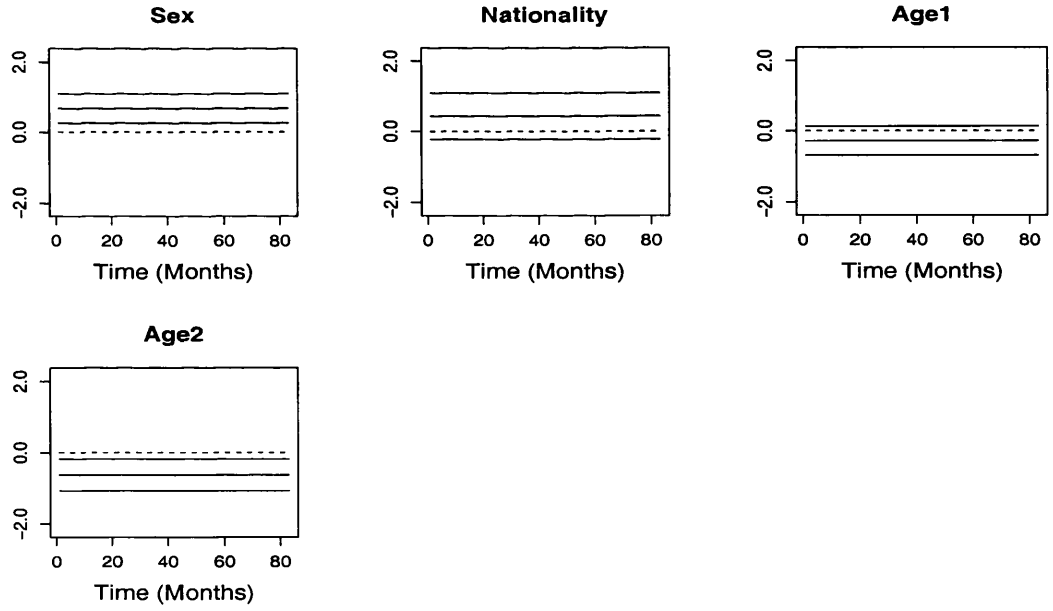


Figure 6.6: *Smooth dynamic covariate effects of the East German unemployment data based on the partial likelihood approach. Outcome is full or part time employment. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

6.3 Unemployment Data Summary

Unemployment data was analysed using both the Poisson type approach from Chapter 4 and the partial likelihood approach from Chapter 5. It was found that males were more likely to gain employment than females, native Germans were more likely to gain employment than foreigners and younger people were more likely to gain employment than those aged over 55 years. Younger people were no more likely to gain employment than middle-aged people. Each of these effects remained constant over time. Interestingly, although those in West Germany were more likely to gain employment than those in East Germany this effect lasted only 20 months. After this time period, there appeared to be no significant differences between the two groups. The estimated baseline hazard from the Poisson type approach was found to increase over time.

The data was divided into two groups to investigate differences between West and East Germany. In West Germany only nationality was found to be significant with native Germans more likely to gain employment than foreigners. In East Germany both sex and age1 were found to be significant with males more likely to gain employment than females and younger individuals more likely to gain employment than those over 55 years old. None of these covariate effects were found to vary over time. For West Germany the baseline hazard was found to be constant over time while in East Germany the baseline hazard increased slightly with time.

Both approaches were found to be very similar. The Poisson type approach has the advantage that the baseline hazard is estimated together with covariate effects, however this approach is slower computationally than the partial likeli-

hood approach and breaks down when dealing with large survival data sets. The partial likelihood approach has the advantage that it can easily deal with large data sets. In the next Section, the partial likelihood approach is used to examine smooth dynamic covariate effects in a large data set from the West of Scotland Coronary Prevention Study.

6.4 West of Scotland Coronary Prevention Study

The West of Scotland Coronary Prevention Study (WOSCOPS) was a primary prevention trial involving 6,595 male participants aged between 45 – 64 years old with raised plasma cholesterol levels. The main aim of the study was to test the hypothesis that reduction of serum cholesterol by treatment with pravastatin over an average period of 5 years would lead to a reduction in fatal and non-fatal myocardial infarction (heart attack). Pravastatin blocks a key step in the body's production of cholesterol and is used to lower cholesterol levels in people with high cholesterol. An overview of the study design is given by The West of Scotland Coronary Prevention Study Group (1992).

Seven endpoints of the trial have been identified as detailed in Table 6.5.

Table 6.5: *Endpoints of the WOSCOP study*

Category	Endpoint
1	Definite coronary heart disease (CHD) death
2	Definite non-fatal myocardial infarction
3	Suspect coronary heart disease (CHD) death
4	Suspect non-fatal myocardial infarction
5	Other cardiac death
6	Other vascular death
7	Other death

Consider 3 outcomes of interest. The first is definite or suspect coronary heart disease (CHD) death (patients fall into category 1 or 3). The second is cardiovascular deaths (patients fall into categories 1, 3, 5 or 6), and the third is all cause mortality (patients fall into categories 1, 3, 5, 6 or 7). A number of baseline

variables were recorded at the beginning of the study. Here, only those variables which have previously been shown to be multivariate predictors of one of the 3 outcomes are considered (see The West of Scotland Coronary Prevention Study Group, 1997). A summary of the variables (with breakdown %'s shown in parenthesis for categorical variables and median values shown for continuous variables) are listed in Table 6.6.

Table 6.6: *Variables of the WOSCOP study*

Categorical	Description (%)	
Treatment	1. Pravastatin (50)	0: Placebo (50)
Current smoker	1: Yes (44)	0: No (56)
Diabetes mellitus	1: Yes (1.2)	0: No (98.8)
Nitrate consumption	1: Yes (2.1)	0: No (97.9)
ECG abnormality	1: Yes (8.1)	0: No (91.9)
Widowed	1: Yes (2.5)	0: No (97.5)
No school leaving cert.	1: Yes (56)	0: No (44)
Continuous	Description (median)	
Age	Age at randomisation (55.2 years)	
SBP	Systolic blood pressure (134 mm Hg)	
DBP	Diastolic blood pressure (84 mm Hg)	

6.4.1 Definite or Suspect CHD Death

Consider the first outcome of interest which is definite or suspect coronary heart disease (CHD) death (total number of events = 102). The independent predictors of definite or suspect CHD death found by The West of Scotland Coronary Prevention Study Group (1997) were treatment allocation, smoking, presence of

diabetes, nitrate consumption, minor electrocardiographic (ECG) abnormality, widowhood, increased age and diastolic blood pressure. Both age and diastolic blood pressure are continuous variables while the others are categorical. In the dynamic hazard model, age is included as a categorical variable such that it is coded as 1 if the participant is older than 55 on entry to the study and coded 0 otherwise. Table 6.7 gives a summary of the output from the Cox proportional hazards model. Multivariate predictors of definite or suspect CHD death are shown in the first column. Given in the second and third columns are estimates of covariate effects β, p Values, risk ratios and 95% confidence intervals. The risk factors for categorical variables correspond to the risk in participants possessing the factor compared to those without it. For continuous risk factors, the risk ratios correspond to changes in the hazard associated with an increase in a specified number of units (shown in parenthesis). Hence the risk ratio for increasing age is given for increments of 5 years while the risk ratio for increasing diastolic blood pressure is given for increments of 10 mm Hg. The second column corresponds to the output given by The West of Scotland Coronary Prevention Study Group (1997) for definite or suspect CHD deaths when age is continuous. The third column gives similar output but with age as a categorical variable.

With age as a categorical variable, the risk ratio for treatment, for example, is 0.66, conditional on the other risk factors in the model. Hence pravastatin therapy reduces the adjusted relative risk of definite or suspect CHD death by 34% ($p = 0.037$). All other variables are significant hence smoking, having diabetes mellitus, consuming nitrates, ECG abnormalities and widowhood all lead to increased risk of definite or suspect CHD death. Increasing age also leads to an increased risk of definite or suspect CHD death and risk increases by around

Table 6.7: *A Cox PH model fitted to the WOSCOPS data with definite or suspect CHD death as outcome*

	Age is Continuous			Age is Categorical		
Variables	β	p Value	Risk Ratio	β	p Value	Risk Ratio
CATEGORICAL						
Treatment	−0.46	0.035	0.65[0.44, 0.97]	−0.42	0.037	0.66[0.44, 0.98]
Current smoker	0.75	0.0002	2.12[1.42, 3.16]	0.71	0.0004	2.04[1.37, 3.04]
Diabetes mellitus	1.12	0.016	3.07[1.24, 7.62]	1.20	0.009	3.33[1.34, 8.27]
Nitrate consump.	1.30	0.0001	3.66[1.94, 6.90]	1.34	< 0.0001	3.82[2.03, 7.21]
ECG abnormality	0.86	0.0004	2.37[1.47, 3.82]	0.89	0.0003	2.44[1.51, 3.95]
Widowed	0.78	0.028	2.18[1.09, 4.39]	0.89	0.012	2.45[1.22, 4.91]
Age				0.79	0.0005	2.20[1.41, 3.43]
CONTINUOUS						
Age (5 years)	0.48	< 0.0001	1.61[1.32, 1.97]			
DBP (10 mm Hg)	0.25	0.010	1.28[1.06, 1.55]	0.24	0.014	1.27[1.05, 1.54]

27% for every 10 unit increase in diastolic blood pressure. This output is based on the assumption of proportional hazards. In standard tests of significance assessing non-proportionality, only the effect of nitrate consumption has significant departures from proportionality ($p = 0.028$). Hence it is likely that the effect of nitrate consumption will vary with time.

Now consider the dynamic hazard model which allows covariate effects to vary with time. The resulting fits obtained from the partial likelihood approach as described in Chapter 5 are shown in Figure 6.7. This plot shows the dynamic effect $\beta(t)$ of each covariate included in the multivariable survival model. Exponentiating $\beta(t)$ gives the risk ratio at each time point. From the plot one can see that those individuals taking pravastatin treatment have a reduced risk of definite or

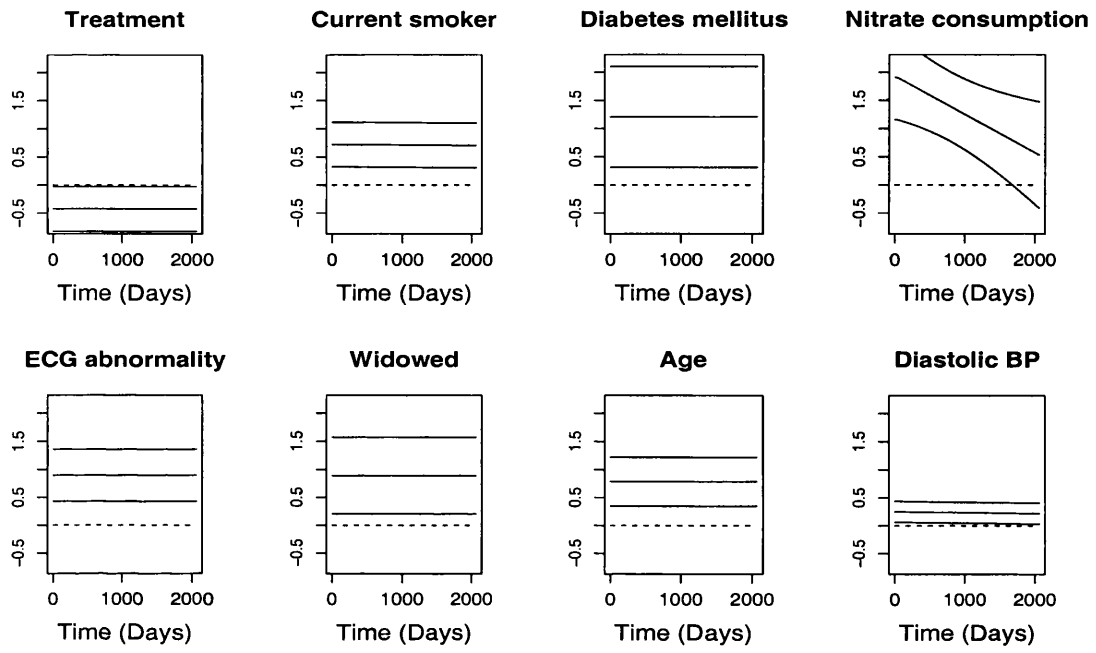


Figure 6.7: *Smooth dynamic covariate effects of the WOSCOPS data with definite or suspect CHD death as outcome. Shown are penalised estimates and* 95% *point-wise confidence intervals. As a reference the zero line is indicated by a dashed line.*

suspect CHD death which remains constant over time. The results are similar for non-smokers and participants without diabetes mellitus. For those taking nitrate medication the risk of definite or suspect CHD death is greater and this effect reduces with time. Those without ECG abnormalities, those who have not been widowed and those who are less than 55 years old all have reduced risk of failure which remains constant over time. Finally a 10 unit increase in diastolic blood pressure leads to an increased risk of definite or suspect CHD death, this effect remains constant over time.

6.4.2 Cardiovascular Deaths

Consider now the second outcome of interest which is cardiovascular deaths (total number of events = 123). The multivariate predictors of cardiovascular deaths were found to be the same as the multivariate predictors of definite or suspect CHD deaths (The West of Scotland Coronary Prevention Study Group, 1997). Table 6.8 shows the predictors of cardiovascular deaths in the first column and estimates of covariate effects β with p Values, risk ratios and confidence intervals in the second and third columns. Again the second column corresponds to age being included as a continuous variable and the third column to age being included in the model as a categorical variable.

Table 6.8: *A Cox PH model fitted to the WOSCOPS data with cardiovascular deaths as outcome*

	Age is Continuous			Age is Categorical		
Variables	β	p Value	Risk Ratio	β	p Value	Risk Ratio
CATEGORICAL						
Treatment	−0.41	0.027	0.67[0.47, 0.96]	−0.43	0.028	0.67[0.47, 0.96]
Current smoker	0.78	< 0.0001	2.22[1.54, 3.21]	0.77	< 0.0001	2.16[1.50, 3.11]
Diabetes mellitus	0.95	0.040	2.59[1.05, 6.38]	1.03	0.025	2.80[1.14, 6.91]
Nitrate consump.	1.21	0.0001	3.35[1.83, 6.13]	1.24	0.0001	3.46[1.89, 6.32]
ECG abnormality	0.80	0.0004	2.22[1.42, 3.46]	0.82	0.0003	2.27[1.46, 3.55]
Widowed	0.71	0.033	2.05[1.06, 3.97]	0.81	0.015	2.26[1.17, 4.35]
Age				0.81	0.0001	2.24[1.50, 3.35]
CONTINUOUS						
Age (5 years)	0.45	< 0.0001	1.57[1.31, 1.87]			
DBP (10 mm Hg)	0.26	0.003	1.30[1.09, 1.54]	0.25	0.004	1.29[1.08, 1.54]

Under the assumption of proportional hazards, when age is included as a categorical variable, treatment with pravastatin reduces the risk of cardiovascular deaths by 33% ($p = 0.028$) conditional on the other risk factors in the model. Additionally smoking, diabetes mellitus, nitrate consumption, minor ECG abnormalities, widowhood and increased age all lead to an increased risk of cardiovascular death. An increase in 10 units of diastolic blood pressure leads to an increased risk of cardiovascular death of around 29%. In standard tests of significance assessing non-proportionality, the assumption of proportional hazards seems reasonable for all covariates, however nitrate consumption is bordering on statistical significance ($p = 0.056$). It is possible therefore that the effect of nitrate consumption may vary with time.

The resulting fits obtained, using the partial likelihood approach, are shown in Figure 6.8. Individuals receiving pravastatin treatment have a reduced risk of cardiovascular death which remains constant over time. Non-smokers and participants without diabetes mellitus have an increased risk of cardiovascular death which remains constant over time. Those on nitrate medication have an increased risk of failure which appears to reduce slightly over time. Those without ECG abnormalities, those who have not been widowed and those in the younger age group all have reduced risk of failure which remains constant over time. An increase of 10 units in diastolic blood pressure leads to an increase in risk of cardiovascular deaths which remains constant with time.

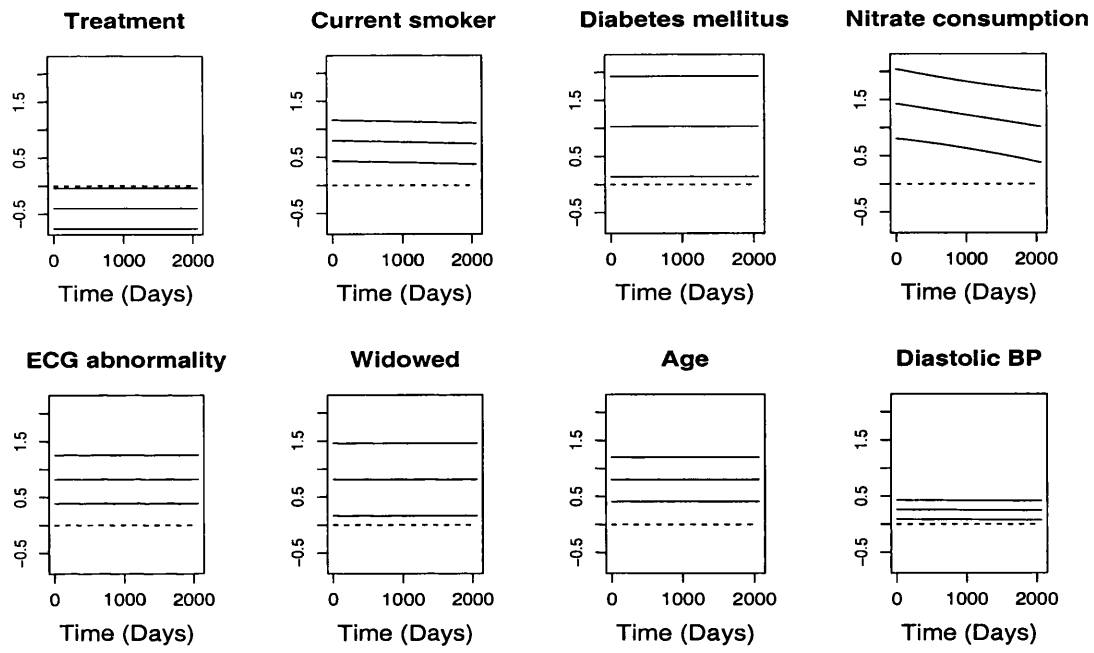


Figure 6.8: *Smooth dynamic covariate effects of the WOSCOPS data with cardiovascular deaths as outcome. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

6.4.3 All Cause Mortality

Finally, consider the third outcome of interest which is all cause mortality (total number of events = 241). The multivariate predictors of all cause mortality differed slightly to the multivariate predictors of definite or suspect CHD deaths and cardiovascular deaths. They were treatment, history of smoking, consumption of nitrates, minor electrocardiographic (ECG) abnormality, widowhood, low educational achievement, age and systolic blood pressure. Both age and systolic blood pressure are continuous variables while the others are categorical. The risk ratio for increasing age is once again given in increments of 5 years while the risk ratio for increasing systolic blood pressure is given in increments of 20 mm Hg. Column 1 of Table 6.9 lists the multivariate predictors of all cause mortality. The second and third columns give output from a fitted Cox proportional haz-

ards model which includes age as a continuous variable (column 2) and age as a categorical variable (column 3).

Table 6.9: *A Cox PH model fitted to the WOSCOPS data with all cause mortality as outcome*

	Age is Continuous			Age is Categorical		
Variables	β	p Value	Risk Ratio	β	p Value	Risk Ratio
CATEGORICAL						
Treatment	−0.27	0.037	0.76[0.59, 0.98]	−0.27	0.036	0.76[0.59, 0.98]
Current smoker	0.81	< 0.0001	2.25[1.73, 2.92]	0.79	< 0.0001	2.20[1.69, 2.85]
Nitrate consump.	0.66	0.017	1.94[1.13, 3.34]	0.68	0.014	1.98[1.15, 3.41]
ECG abnormality	0.49	0.006	1.64[1.15, 2.33]	0.50	0.006	1.65[1.16, 2.34]
Widowed	0.69	0.0053	2.00[1.23, 3.26]	0.77	0.002	2.15[1.32, 3.50]
No school cert.	0.31	0.024	1.37[1.04, 1.80]	0.33	0.018	1.39[1.06, 1.83]
Age				0.85	< 0.0001	2.34[1.74, 3.14]
CONTINUOUS						
Age (5 years)	0.43	< 0.0001	1.54[1.35, 1.75]			
SBP (20 mm Hg)	0.24	0.0008	1.27[1.10, 1.46]	0.26	0.0002	1.30[1.13, 1.49]

Thus when age is included as a categorical variable, treatment with pravastatin reduces the risk of all cause mortality by 24% ($p = 0.036$) conditional on all other risk factors. Additionally smoking, consuming nitrates, ECG abnormalities, widowhood, low educational achievement and increased age all lead to increased risk of all cause mortality. An increase of 20 units in systolic blood pressure leads to an increased risk of cardiovascular death of around 30%. In standard tests of significance assessing non-proportionality, only the effect of nitrate consumption has significant departures from proportionality ($p = 0.021$). The effect of ECG abnormality is bordering on statistical significance ($p = 0.053$). Hence it is likely that the effect of nitrate consumption will vary with time and the effect of ECG

abnormality may also vary over time.

Smooth fits are obtained using the partial likelihood approach as shown in Figure 6.9. Participants receiving pravastatin treatment have a reduced risk of

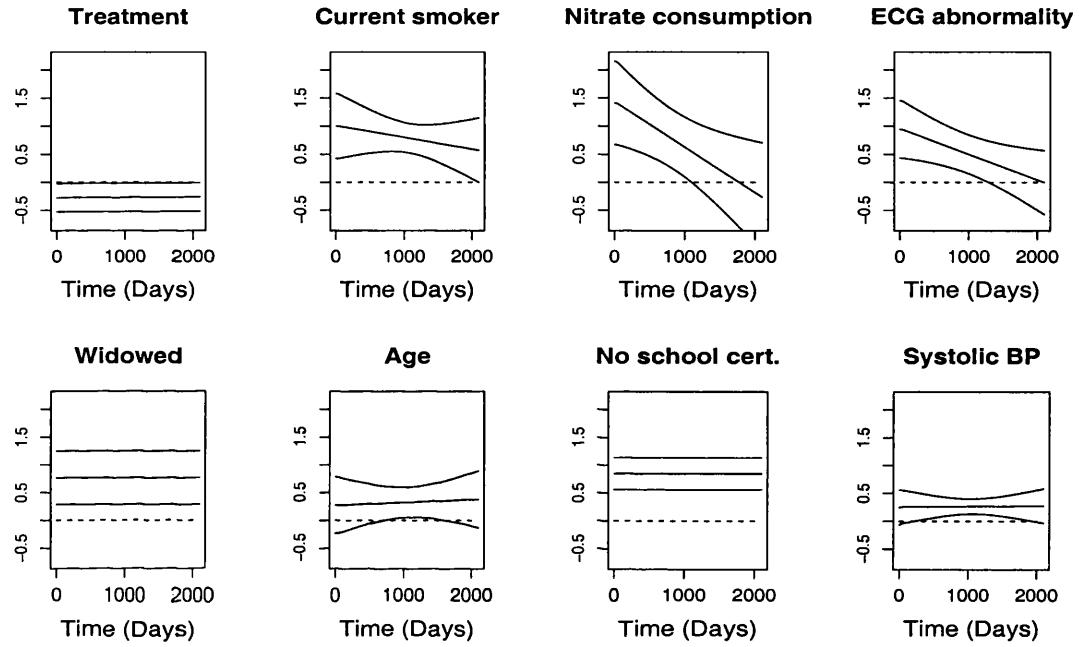


Figure 6.9: *Smooth dynamic covariate effects of the WOSCOPS data with all cause mortality as outcome. Shown are penalised estimates and* 95% *pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

all cause mortality which remains constant over time. Smokers have an increased risk of death which appears to reduce slightly over time. For those on nitrate medication and those with minor ECG abnormalities the risk of death is increased, however, both these effects appear to reduce with time. Those who have been widowed, those with low educational achievement and those who are over 55 years old have an increased risk of death which remains constant over time.

6.4.4 All Cause Mortality – 10 year follow-up

Participants' survival was followed-up for a further 5 year period after the close of the WOSCOP study. As a result, additional information regarding survival time and status was made available for some of the endpoints, including all cause mortality. This enables investigation of risk factors on all cause mortality over an average period of 10 years (from the beginning of the WOSCOP study until 19^{th} March 2000). A Cox proportional hazards model fitted to the data shows that covariates found to be predictors of all cause mortality remain significant over this 10 year period. Smooth covariate effects for the 10 year all cause mortality

Table 6.10: *A Cox PH model fitted to the WOSCOPS data with all cause mortality (10 year follow-up) as outcome*

	Age is Continuous			Age is Categorical		
Variables	β	p Value	Risk Ratio	β	p Value	Risk Ratio
CATEGORICAL						
Treatment	−0.19	0.012	0.83[0.72, 0.96]	−0.19	0.012	0.83[0.71, 0.96]
Current smoker	0.70	< 0.0001	2.02[1.73, 2.35]	0.68	< 0.0001	1.97[1.69, 2.29]
Nitrate consump.	0.73	< 0.0001	2.08[1.51, 2.87]	0.76	< 0.0001	2.15[1.56, 2.96]
ECG abnormality	0.33	0.003	1.40[1.12, 1.74]	0.34	0.0023	1.41[1.13, 1.76]
Widowed	0.43	0.0092	1.54[1.13, 2.14]	0.52	0.0017	1.68[1.22, 2.33]
No school cert.	0.26	0.001	1.30[1.11, 1.52]	0.28	0.0005	1.32[1.13, 1.55]
Age				0.83	< 0.0001	2.29[1.93, 2.71]
CONTINUOUS						
Age (5 years)	0.43	< 0.0001	1.54[1.40, 1.66]			
SBP (20 mm Hg)	0.20	< 0.0001	1.22[1.13, 1.31]	0.22	< 0.0001	1.24[1.15, 1.37]

data are now obtained using the partial likelihood approach as shown in Figure 6.10.

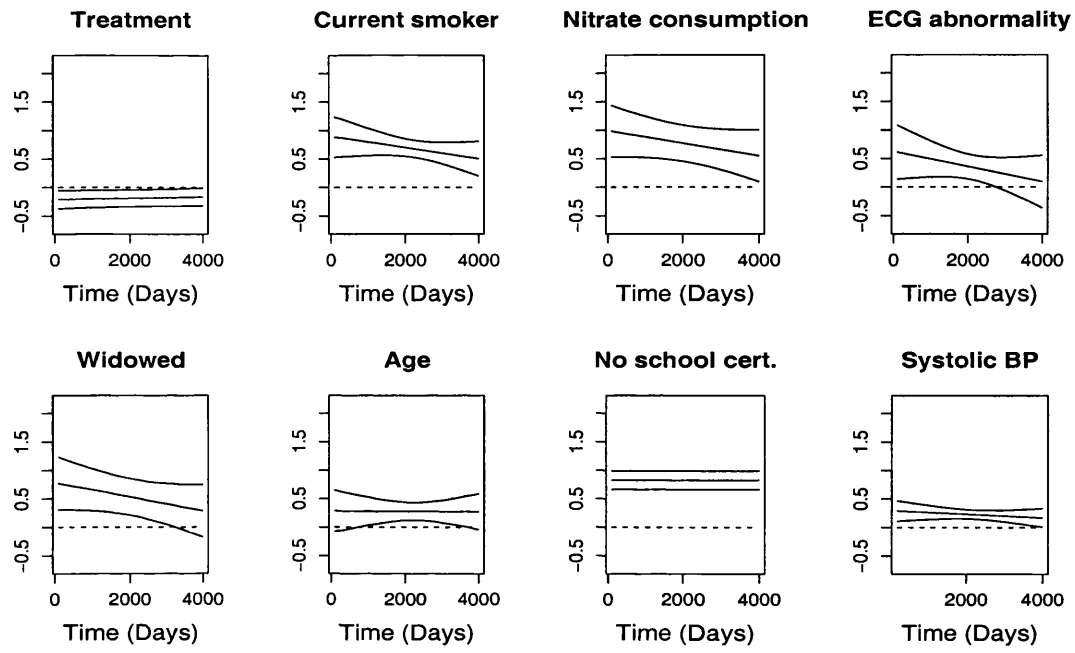


Figure 6.10: *Smooth dynamic covariate effects of the WOSCOPS data with all cause mortality (10 year follow-up) as outcome. Shown are penalised estimates and 95% pointwise confidence intervals. As a reference the zero line is indicated by a dashed line.*

Participants receiving pravastatin treatment have a reduced risk of all cause mortality which remains constant over time. Smokers have an increased risk of death which appears to reduce slightly with time. For those on nitrate medication there is an increased risk of death which appears to reduce over time. Those with minor ECG abnormalities have an increased risk of death although this effect also reduces with time. Those who have not been widowed have a reduced risk of death which reduces with time. Those over 55 have an increased risk which is constant over time, as do those with low educational achievement. Finally a 20 unit increase of systolic blood pressure leads to an increase in risk of death for which there appears to be a slight reduction with time.

6.5 WOSCOPS Data Summary

The West of Scotland Coronary Prevention Study achieved its aim of demonstrating the benefit of pravastatin therapy in the prevention of coronary heart disease events in men with high cholesterol levels. During the study multiple endpoints were identified and three of those endpoints were considered here. These were definite or suspect CHD death, cardiovascular deaths and all cause mortality. In each case the benefit of pravastatin therapy was clear. As well as the treatment group, other baseline variables were taken into consideration and multivariate predictors of each endpoint were analysed using the Cox proportional hazards model. This model assumes that the effects of covariates are constant. By allowing the covariate effects to vary with time this strict assumption was avoided. For definite or suspect CHD deaths, nitrate consumption was found to be the only effect which varied with time. The assumption of proportional hazards seemed reasonable for each of the other covariates in the model. A similar conclusion was reached for those with cardiovascular death as an endpoint. For all cause mortality, the effects of both nitrate consumption and minor ECG abnormality changed with time. The effect of smoking also appeared to be time dynamic. For the remaining variables the assumption of proportionality seemed reasonable. For all cause mortality 10 year follow-up the effects of smoking, ECG abnormality and nitrate consumption appeared to reduce over time. The effect of systolic blood pressure also appeared to reduce slightly over the 10 year period. Finally, the effect of being widowed showed evidence of reducing over a 10 year follow-up period.

The effect of nitrate consumption is interesting as it reduces with time for

each of the endpoints considered. Nitrates are useful in relieving angina pain and in preventing angina in the long term. It is plausible that the group of subjects taking nitrates are of heterogeneous risk. Subjects recently diagnosed with angina are known to be at very high risk while others may have been misdiagnosed with angina and incorrectly treated with nitrates or are patients in whom nitrates are being used as a diagnostic tool to see if the treatment prevents effort related chest pain. If this is true then the highest risk patients will leave the risk set first creating a reducing hazard as the group taking nitrates become dominated by low risk or misdiagnosed patients.

Chapter 7

Conclusions and Discussion

The aim of this thesis was to investigate dynamic covariate effects in survival data. Based on an extension of the Cox proportional hazards model this research investigated two different approaches to the estimation of dynamic covariate effects. One approach, based on the likelihood function, had the advantage that it allows covariate effects to be estimated together with the baseline hazard function while the other approach, based on the partial likelihood function, was numerically faster and allowed for estimation of dynamic covariate effects in large data sets. The main novelty of this work was the link to linear mixed models, which is used in a hybrid form for smoothing parameter selection. This routine was used for both survival models fitted using a Poisson tye approach and for survival models fitted using the partial likelihood function. The findings from this research have been significant in that they have shown how new insight into data sets can be gained by allowing covariate effects to vary with time. As well as applying these approaches to new data sets, across a range of fields, it may also be of interest to look at data sets which have previously been analysed using the Cox proportional

hazards model to see if additional insight into the effects of covariates may be obtained, particularly for those studies with a long follow-up period.

Chapter 2 gave an introduction to survival analysis and focused on the Cox proportional hazards model which is used to explore the relationship between covariates and survival time. The underlying baseline hazard function in the Cox model is estimated non-parametrically while the partial likelihood function is used to estimate covariate effects. The Cox model assumes proportionality of the hazards which means that covariate effects are assumed to be constant over time. This assumption may be questionable especially in the case of long-term follow up, hence a dynamic model allowing covariate effects to change with time was suggested.

Dynamic covariate effects were estimated smoothly over time using penalised (P-spline) smoothing. Chapter 3 compares this method to other smoothing methods and motivates the use of P-splines in smooth hazard modelling. A link between P-splines and generalised linear mixed models is utilised to enable data driven estimates of smoothing parameters controlling the amount of smoothing.

Chapter 4 introduced the first of two approaches to the smooth estimation of dynamic covariate effects. An approach based on the likelihood function was considered, which was an extention of the work carried out in recent years by Cai et al. (2002) and Cai and Betensky (2003). However, the advantage of this method is that both dynamic covariates effects and a smooth baseline hazard are modelled simultaneously. Non-proportional hazard functions were fitted using Poisson regression resulting from numerical integration of the cumulative hazard function. Smooth estimation was carried out via P-spline smoothing. The

connection between P-splines and generalised linear mixed models was used to choose appropriate smoothing parameters and the approach was evaluated with simulations.

A second approach to the smooth estimation of dynamic covariate effects was introduced in Chapter 5. This approach follows on from the work by several authors including Gray (1994). There are noticeable differences, however, in the work produced here and the work of Gray. Gray's main focus is on testing while this work focuses more on modelling. Furthermore, Gray (1994) leaves the data driven choice of smoothing parameter aside while this work treats this choice explicitly. Finally the paper by Gray has no links to mixed models which is emphasised here. Based on the partial likelihood function, this approach was numerically faster than the Poisson type approach from Chapter 4 and therefore allowed estimation of dynamic effects in larger simulated data sets. Dynamic covariate effects were estimated using P-splines with the link between P-splines and generalised linear mixed models utilised for smoothing parameter selection. Simulations demonstrated the performance of this approach.

In Chapter 6 both the Poisson type approach from Chapter 4 and partial likelihood approach from Chapter 5 were applied to unemployment data from the German Socio-Economic Panel. Full time or part time employment was considered as endpoint. An interesting dynamic effect was found while using both approaches. Those in West Germany were found to have a greater chance of gaining employment than those in East Germany, however this effect only remained significant for the first 20 months. After this time the effect became non-significant. Differences within West and East Germany were then examined with no dynamic effects found. Further analysis of the unemployment data may

include considering individuals who, had their first spell not been completely observed, were now in their second spell of unemployment.

The partial likelihood approach was also applied to data from the West of Scotland Coronary Prevention Study (WOSCOPS). The endpoints considered were definite or suspect CHD death, cardiovascular deaths and all cause mortality. Only those covariates which had previously been found to be significant by the WOSCOP study group were included in the model for each endpoint. For definite or suspect CHD deaths and for cardiovascular deaths nitrate consumption was the only effect found to vary with time. For all cause mortality, the effects of both nitrate consumption and minor ECG abnormality varied with time while smoking also showed some dynamic effects. A further 5 year follow-up period allowed the effects of all cause mortality to be investigated over an average period of 10 years. The multivariate predictors of all cause mortality remained significant over this 10 year period and the effects of smoking, ECG abnormality, nitrate consumption and widowhood all reduced over time. The effect of systolic blood pressure also reduced slightly with time. As an extension to the West of Scotland Coronary Prevention Study more endpoints could be considered. It would also be of interest to have 10 year follow-up data made available for the outcomes definite or suspect CHD death and cardiovascular deaths in order to see if effects that are constant over 5 years remain so over this longer follow-up period. Note that in this analysis the effects of competing risks have not been considered. However in this data set the event rates are very low and this is unlikely to cause any practical problems.

So far only right censored survival data has been considered, however, it would be of interest to extend both approaches to model left or interval censored data. It would also be of interest to model data with multiple events. This is possible

under the framework of the Poisson type approach since pseudo observations Y_{ik} indicate with a one within which grid point an event has been experienced. This framework also allows for modelling complicated risk patterns such as staggered entry where individuals may begin and end their period of risk at different times. Here, an individual is only at risk within certain grid points. Further work on both approaches includes extended modelling to allow for the effects of frailties. A final point to note is that models which do not require the assumption of proportional hazards have not been considered in these analyses. These include the proportional odds model and the accelerated failure time model where explanatory variables measured on an individual are assumed to act multiplicatively with time.

Much of the time in this research has been spent writing programs for both approaches, however, there is still scope for improvement, particularly for the Poisson type approach. For one data set simulated with a constant covariate effect and constant baseline hazard ($N = 400$, drop out probability of 3% at each time interval t to $t+1$) a data driven estimate is obtained in around 1 and a half minutes using the Poisson type approach compared to around 8 seconds using the partial likelihood approach (around 35 seconds for the partial likelihood approach when $N = 4000$). Timings are based on running the programs on two 1.2 GHz Sparc processors, with R version 2.0.0, under the Solaris Unix 2.8 operating system. The main cause of the difference in computation time appears to be due to having to sum over up to q integration points for each individual when using the Poisson type approach. For $N = 400$ the number of knots hardly affects the speed of computation for the partial likelihood approach (a difference of less than 1 second when going from 20 knots to 30 knots). For the Poisson type approach

however, increasing the number of knots from 20 to 30 leads to an increase in the computing time required of around 30%. If improvements to the coding of both programs were made, then this would enable more samples, of varying sample sizes and censoring rates, to be generated in the simulation study allowing more comparisons between the two approaches. One could also more easily investigate the effects of using a different spline basis, varying the degree of splines used and using more or less knots, although this has already been investigated in, for example, the book by Ruppert et al. (2003).

For the partial likelihood approach the cumulative hazard function was allowed to factorise to the covariate effects multiplied by the cumulative baseline hazard function, by pretending that $\beta(t) = \beta$. The cumulative baseline hazard was then left unspecified while $\beta(t)$ was estimated smoothly. In contrast, with the Poisson type approach, factorisation of the cumulative hazard did not exist. Integration of the cumulative hazard function was achieved using numerical integration. Comparisons between the two approaches were made by running simulation studies and by observing the fit obtained by applying both approaches to the unemployment data from the German Socio-Economic Panel. In terms of further work it would be of interest to be able to measure the differences between the two approaches numerically.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* pp. 267–281.

Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function, *Journal of the Royal Statistical Society, Series B* **42**: 322–327.

Boneva, L. I., Kendall, D. G. and Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the data analyst (with discussion), *Journal of the Royal Statistical Society, Series B* **33**: 1–70.

Booth, J. G. and Hobert, J. H. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society, Series B* **62**: 265–285.

Bowman, A. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis*, Oxford University Press.

Breslow, N. E. (1972). Comment on D. R. Cox (1972) paper, *Journal of the Royal Statistical Society, Series B* **34**: 216–217.

Breslow, N. E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**: 89–100.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**: 81–91.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval censored data with penalized spline, *Biometrics* **59**: 570–579.

Cai, T., Hyndman, R. J. and Wand, M. P. (2002). Mixed model-based hazard estimation, *Journal of Computational and Graphical Statistics* **11**: 784–798.

Cai, Z. and Sun, Y. (2002). Local linear estimation for time-dependent coefficients in Cox's regression models, *Scandinavian Journal of Statistics* **30**: 93–111.

Collett, D. (1994). *Modelling Survival Data in Medical Research*, Chapman and Hall, London.

Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**: 187–220.

Cox, D. R. (1975). Partial likelihood, *Biometrika* **62**: 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by method of generalized cross validation, *Numerische Mathematik* **31**: 377–403.

de Boor, C. (1972). On calculating with B-splines, *Journal of Approximation Theory* **6**: 50–62.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag, New York.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association* **72**: 557–565.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science* **11**: 89–121.

Eilers, P. H. C. and Marx, B. D. (2004). Splines, knots, and penalties. Preprint.

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression, *The Annals of Statistics* **25**: 1661–1690.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples, *Biometrika* **72**: 557–565.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika* **58**: 255–277.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**: 515–526.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis, *Journal of the American Statistical Association* **87**: 942–951.

Gray, R. J. (1994). Spline-based tests in survival analysis, *Biometrics* **50**: 640–652.

Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Skandinavisk Aktuarietidskrift* pp. 119–134.

Hastie, T. (1996). Pseudosplines, *Journal of the Royal Statistical Society, Series B* **58**: 379–396.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models, *Journal of the Royal Statistical Society, Series B* **55**: 757–796.

Hess, K. R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression, *Statistics in Medicine* **14**: 1707–1723.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model, *Biometrika* **60**: 267–278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**: 457–481.

Kauermann, G. (2003). Some formulae on P-spline smoothing in generalized response models, *Technical report*, University of Bielefeld.

Kauermann, G. (2004). A note on smoothing parameter selection for penalized spline smoothing, *Journal of Statistical Planning and Inference* **127**: 53–69.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation, *Journal of the American Statistical Association* **96**: 1387–1396.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Kooperberg, C., Stone, C. J. and Troung, Y. K. (1995). Hazard regression, *Journal of the American Statistical Association* **90**: 78–94.

Mallows, C. L. (1973). Some comments on C_p, *Technometrics* **15**: 661–675.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**: 719–748.

Nadaraya, E. A. (1964). On estimating regression, *Theory of Probability and its Applications* **10**: 186–190.

Oakes, D. (1972). Comment on D. R. Cox (1972) paper, *Journal of the Royal Statistical Society, Series B* **34**: 208.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion), *Statistical Science* **1**: 502–527.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation, *SIAM Journal on Scientific and Statistical Computing* **9**: 531–542.

Parmar, K. B. and Machin, D. (1995). *Survival Analysis: A Practical Approach*, Wiley, Chichester, UK.

Reinsch, C. H. (1967). Smoothing by spline functions, *Numerische Mathematik* **10**: 177–183.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines, *Biometrics* **51**: 874–887.

Ruppert, D. (2002). Selecting the number of knots for penalized splines, *Journal of Computational and Graphical Statistics* **11**: 735–757.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, UK.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*, KTK Scientific Publishers, Tokyo.

Sasieni, P. (1999). Cox regression model, *in* P. Armitage and T. Colton (eds), *Encyclopedia of Biostatistics*, Vol. 1, Wiley, New York, pp. 1006–1020.

Schall, R. (1991). Estimation in generalized linear models with random effects, *Biometrika* **78**: 719–727.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions, *Quarterly Applied Maths* **4**: 45–99.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high-dimensional integrals, *Journal of the Royal Statistical Society, Series B* **53**: 749–760.

Stone, C. J. (1986). Comment: Generalized additive models, *Journal of the Royal Statistical Society, Series B* **1**: 312–314.

Strawderman, R. L. and Tsiatis, A. A. (1996). On the asymptotic properties of a flexible hazard estimator, *The Annals of Statistics* **24**: 41–63.

The West of Scotland Coronary Prevention Study Group (1992). A coronary primary prevention study of Scottish men aged 45–64 years: trial design, *Journal of Clinical Epidemiology* **45**: 849–860.

The West of Scotland Coronary Prevention Study Group (1997). Baseline risk factors and their association with outcome in the West of Scotland Coronary Prevention Study, *American Journal of Cardiology* **79**: 756–762.

Wand, M. P. (2003). Smoothing and mixed models, *Computational Statistics* **18**: 223–249.

Watson, G. S. (1964). Smooth regression analysis, *Sankhya: The Indian Journal of Statistics, Series A* **26**: 359–372.

Whittaker, E. T. (1923). On a new method of graduation, *Proceedings of the Edinburgh Mathematical Society* **41**: 63–75.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society, Series B* **62**: 413–428.

Wood, S. N. (2003). Thin-plate regression splines, *Journal of the Royal Statistical Society, Series B* **65**: 95–114.

Wood, S. N. (2004). On confidence intervals for GAMs based on penalized regression splines, *Technical report*, University of Glasgow.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach, *The Annals of Statistics* **18**: 329–253.